Unraveling transcript-based variability of host responses to Tuberculosis

Dissertation

zur Erlangung des akademischen Grades

Doctor of Philosophy
(Ph. D.)

im Fach Biologie

eingereicht an der

Lebenswissenschaftlichen Fakultät

der Humboldt-Universität zu Berlin

von

Teresa Domaszewska (M. Sc.)

Durchgeführt am Max-Planck-Institut für Infektionsbiologie in der Abteilung Immunologie

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst
Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Bernhard Grimm

Gutachter/innen: 1. Stefan H. E. Kaufmann
2. Barbara Broeker
3. Arturo Zychlinsky

Tag der mündlichen Prüfung: 10.01.2019

DECLARATION

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids.
I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree.
I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected.
I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on 5 th March 2015.
Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

The work presented here has been performed under direct supervision of Dr. January Weiner (Max Planck Institute for Infection Biology, Department of Immunology), to whom I am immensely thankful for guiding me during the years of my doctoral studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AUC – area under curve

BCG – *Bacillus Calmette-Guérin*

bp – base pair

BTM – blood transcriptional module

CART – classification and regression tree

Cy3 – cyanine-3

Cy5 – cyanine-5

DC – dendritic cell

DNA – deoxyribonucleic acid

DRS – disease risk score

ELISA – enzyme-linked immunosorbent assay

FcGR – Fc γ receptor

GEO – Gene Expression Omnibus Database

GLM – generalized linear model,

GO – gene ontology

GSEA – Gene Set Enrichment Analysis

HC – healthy control

HIV – human immunodeficiency virus

HIV+ – infected with human immunodeficiency virus

IFN – interferon

IFNAR – interferon-α/β receptor

IFNGR – interferon-γ receptor

IGRA – interferon-γ release assay

IL – interleukin

ILR – interleukin receptor

IQR – interquartile range

LAM – lipopolysaccharide lipoarabinomannan

log_2FC – base 2 logarithm of fold change

LTBI – latent tuberculosis infection

Maf – *Mycobacterium africanum*

MDS – meta-dataset

MHC – major histocompatibility complex

ml – mililiter

ML – machine learning

mRNA – messenger ribonucleic acid

MSigDB – Molecular Signatures Database

Mtb – *Mycobacterium tuberculosis*

NAAT – nucleic acid amplification test

NK – natural killer cell

OD – other disease

PBMC – peripheral blood mononuclear cell

PC – principal component

PCA – principal component analysis

PCR – polymerase chain reaction

PD-L1 – programmed death ligand 1

PD-L2 – programmed death ligand 2

PET-CT positron emission tomography – computed tomography

p.i. – post infection

PMA – phorbol 12-myristate 13-acetate

PMN – polymorphonuclear cell

PPD – purified protein derivative

RF – random forest

RIN – RNA integrity number

RNA – ribonucleic acid

RNS – reactive nitrogen species

ROC – receiver-operator characteristic

ROS – reactive oxygen species

rRNA – ribosomal ribonucleic acid

RT-PCR – reverse-transcription polymerase chain reaction

SA – South Africa

SAGE – serial analysis of gene expression

SNP – single nucleotide polymorphism

SPF – specific pathogen-free

TB – tuberculosis

TF – transcription factor

TNF – tumor necrosis factor

tRNA – transfer ribonucleic acid

TST – tuberculin skin test

UK – United Kingdom

WB – whole blood

WHO – World Health Organization

ABSTRACT *(English)*

Over 10 million tuberculosis (TB) cases are being reported annually and the World Health Organization (WHO) estimates that up to the 1/3 of the world population is infected with *Mycobacterium tuberculosis* (Mtb). Between 5 and 10% of the latently infected individuals develop TB during their lifetime. Yet, despite over 100 years of research since Mtb has been identified, we are not able to define all the factors which are responsible for the different infection outcomes in the hosts.

In this thesis I investigate the variability in the response to TB presented by different hosts. In one approach, I collect publicly available transcriptomic datasets from TB patients and healthy donors. Using Gene Set Enrichment Analysis (GSEA) I examine transcriptional profiles of individuals with TB. In particular, focus is brought to interferon (IFN) signaling which has been previously described as crucial for the disease outcome. I show that patients lacking IFN signature are present in the studied cohorts and investigate whether these patients present different phenotype than patients with strong regulation of IFN responses. Moreover, by focusing on patients lacking IFN response I try to unearth mechanisms present in all patient groups but dominated by the signal of IFN response. I show that strong regulation of IFN genes is related to severe pathology in the lungs of TB patients and that it is reflected by the levels of IFN-inducible cytokines in blood of healthy volunteers after vaccination with FLUAD® vaccine. Using Machine Learning (ML) methods, I identify and compare transcriptomic signatures of the patients presenting and lacking the IFN response.

In the second approach I study the differences in the transcriptional responses to Mtb infection in human cohorts and two different mouse models. The immunity in infection, inflammation and malignancy differs markedly in man and mouse. Nevertheless, there are elements of immune system which have been conserved between the species. I propose a novel data integration approach which identifies concordant and discordant elements of gene expression regulation in heterologous datasets. The analysis is based on publicly available as well as novel experimental data acquired thanks to collaboration with my colleagues from the Department of Immunology and Microarray Core Facility of Max Planck Institute for Infection Biology (MPIIB). Additionally, I focus on the comparison of human and murine transcriptional responses to TB in whole blood (WB) and in macrophages. The results indicate profound differences between regulation of innate and adaptive immunity in man and mouse upon Mtb infection. I characterize differential regulation of T-cell related genes corresponding to the differences in phenotype between TB high and low susceptible mouse strains and identify the time point of 21 days p.i. of mice as best reflection of transcriptional responses in the studied human cohorts.

The implemented approaches facilitate the choice of an appropriate animal model for studies of the human immune response to a particular disease and provide the basis for better understanding of differences in the outcomes of Mtb infection in individual hosts.

ABSTRACT *(German)*

Jedes Jahr treten weltweit über zehn Millionen Fälle von Tuberkulose (TB) auf. Die Weltgesundheitsorganisation (WHO) schätzt, dass ein Drittel der Weltbevölkerung mit dem Erreger *Mycobacterium tuberculosis* (Mtb) infiziert ist. Bei fünf bis zehn Prozent aller latent Infizierten bricht Tuberkulose im Laufe des Lebens aus. Dennoch sind bereits 100 Jahre seit der Entdeckung von Mtb vergangen, ohne dass die entscheidenden Faktoren für den unterschiedlichen Infektionsverlauf bekannt wären.

In dieser Arbeit untersuche ich die unterschiedlichen Reaktionen auf eine Tuberkuloseinfektion in verschiedenen Wirten. In meinem ersten Ansatz habe ich öffentlich zugängliche Transkriptom-Datensätze von Tuberkulosepatienten und gesunden Probanden ausgewertet. Mit Hilfe der Gensatzanreicherungs-Analyse (eng. Gene Set Enrichment Analysis, GSEA) habe ich die Transkriptionsprofile von Tuberkulosepatienten betrachtet. Das besondere Augenmerk lag hierbei auf der Interferon (IFN)-Signalkaskade, die für den Krankheitsverlauf von besonderer Bedeutung ist. In dieser Arbeit zeige ich zunächst, dass Patienten ohne eine IFN-Signatur in der untersuchten Kohorte vorkommen und widme mich im Anschluss der Frage, ob diese Patienten einen anderen Phänotypus haben als jene mit einer starken IFN-Antwort. Indem ich nur Patienten ohne IFN-Antwort betrachte, werden Mechanismen deutlich, die allen Patientengruppen gemein sind, aber vorher von der starken IFN-Signatur überlagert wurden. Ich belege in dieser Arbeit, dass eine starke IFN-Regulation auch mit einer ausgeprägten Lungenpathologie in Tuberkulosepatienten einhergeht. Passend hierzu weisen auch gesunde Probanden nach Verabreichung des Impfstoffs FLUAD® einen erhöhten Blutwert IFN-induzierter Zytokine auf. Mit Hilfe maschinellen Lernens konnte ich Transkriptomsignaturen der Patienten mit bzw. ohne IFN-Antwort identifizieren und vergleichen.

Im zweiten Ansatz widme ich mich den unterschiedlichen Transkriptionsantworten auf Mtb-Infektionen in humanen Kohorten und zwei verschiedenen Mausmodellen. Der humanen und der murinen Immunantwort auf Infektionen unterliegen gravierende Unterschiede. Trotzdem sind einige Elemente des Immunsystems in beiden Arten konserviert. In dieser Arbeit präsentiere ich einen neuen Ansatz der Datenintegration, der die Identifizierung von übereinstimmenden und nicht übereinstimmenden Regulationselementen der Genexpression in heterogenen Datensätzen ermöglicht. Die Analyse basiert auf öffentlich zugänglichen sowie de-novo-generierten Datensätzen, zu denen ich durch wissenschaftliche Kollaborationen meiner Kollegen in der Abteilung Immunologie sowie der zentralen Einheit Microarray des Max-Planck-Instituts für Infektionsbiologie, Zugang erhalten habe. Des Weiteren liegt ein Schwerpunkt auf der vergleichenden Analyse humaner und muriner Transkriptionsantworten auf Tuberkulose in Vollblut und Makrophagen. Die erhaltenen Ergebnisse weisen auf einen signifikanten Unterschied in der Regulierung der angeborenen sowie der erworbenen Immunität in Mensch und Maus als Reaktion auf eine Mtb-Infektion hin. In dieser Arbeit charakterisiere ich die unterschiedliche Regulierung von T-Zell bezogenen Genen, die mit unterschiedlich ausgeprägten Phänotypen bei stark oder schwach TB-anfälligen Mausstämmen korrespondiert. Darüber hinaus habe ich den 21. Tag nach einer Tuberkuloseinfektion in Mäusen als Zeitpunkt ermittelt, der die Transkriptionsantworten in den untersuchten humanen Kohorten am besten widerspiegelt.

Die angewandten Ansätze erleichtern die Auswahl des am besten geeigneten Tiermodells für die Erforschung der humanen Immunantwort auf eine ausgewählte Krankheit und liefern die Basis für ein besseres Verständnis der unterschiedlichen Krankheitsverläufe in Mtb-infizierten Patienten.

1. CHAPTER 1: INTRODUCTION AND GOALS

Tuberculosis is an airborne infectious disease caused by Mycobacteria, usually *Mycobacterium tuberculosis*. It typically affects the lungs and causes symptoms including fever, weight-loss, night sweats, and chronic cough containing blood-stained sputum (Hopewell, 2017). TB remains a threat to public health with an enormous disease burden of 10.4 million cases and 1.7 million deaths per year, as estimated for 2016 (WHO, 2017). One of the challenges in preventing TB results from the fact that only part (around 10%) of people infected with Mtb progress to clinical disease; however, so far there is no efficient way of predicting who of the infected individuals will develop active TB and therefore should start preventive treatment.

The research presented in this thesis focuses on TB. Understanding of TB in humans and advancing the translation of TB research has been the main motivation behind the performed work. For this reason, in the first section of my thesis I introduce the reader to the topic of TB by reviewing the most important aspects of epidemiology, treatment, detection and prevention of TB as well as the current state of knowledge about immune response against TB focusing on the broadly discussed topic of the role of IFN responses. This is followed by a short discussion of variability of the events succeeding the Mtb infection which starts at the level of cellular events and finds its consequence in the ultimate disease outcome. Since our understanding of infectious disease mechanisms is to a major extent based on experimental murine models, my second point of focus is the description of the advances brought to the field of TB by investigating murine models and the need of finding a method to choose animal models which best approximate particular aspects of human disease. Lastly, I introduce the reader to the high-throughput experimental methods used to generate the transcriptomic datasets analyzed in this thesis as well as computational methods on which I base the presented results.

1.1. TUBERCULOSIS

1.1.1. Epidemiology of TB

TB leads the statistics of infectious diseases caused by a single infectious agent which have a worldwide impact. About 6.3 million new cases and 1.7 million deaths attributed to TB were reported in 2016, and an estimated 4.1 million remain undiagnosed (WHO, 2017). TB is more prevalent in the low-income areas and 56% of the cases reported in 2016 come from only five countries: China, India, Indonesia, Pakistan and the Philippines. The risk of TB is highly increased among the individuals infected with human immunodeficiency virus (HIV) which comprise 20% of the TB cases detected in the 2016. Other risk factors include poverty, smoking and undernutrition (WHO, 2013a). The disease is more frequent among men than among women and affects mostly adults which might be at least in part attributed to the fact that 80% of the newborns are vaccinated against TB (WHO, 2017).

The spreading drug resistance of mycobacterial strains increases the gravity of TB threat. In 2016 the resistance to rifampicin – a first line anti-TB drug – was detected in more than 500,000 patients. The treatment of drug-susceptible TB is costly and long, lasting at least 6 months and giving success rate of around 83%. In comparison, the shortest regimens for drug-resistant TB is twice as long and gives the success rate of around 54% (WHO, 2014, 2017) . Financing TB prevention and treatment is a major challenge and while the funds dedicated to fighting TB have been increasing during the last 10 years, funding gaps still exist and require investments from the side of governments and organizations on both national and international levels (WHO, 2017).

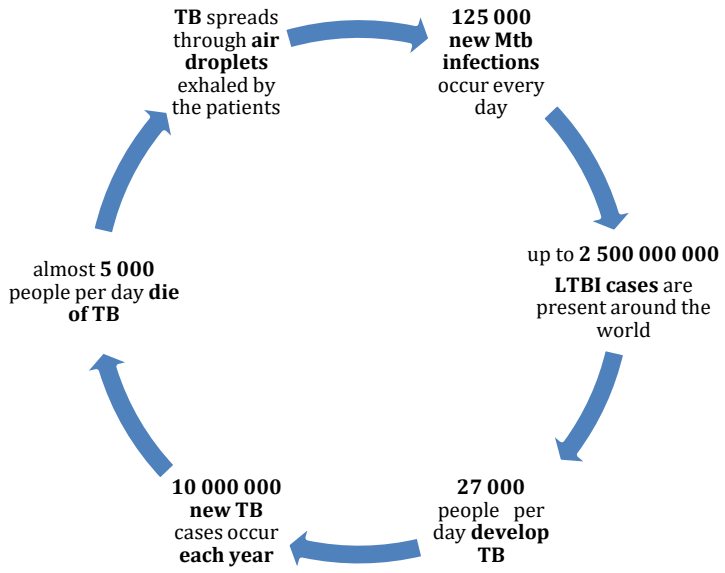


Figure 1 Vicious circle of TB

Around 125,000 new Mtb infections occur each day through spreading of air droplets by the TB patients. This adds up to over 2 million latent TB infections (LTBI) around the world according to the estimations of WHO (2017). 27,000 of the infected people daily develop TB, which gives rise to around 10 million new TB cases per year and results in 5,000 TB deaths daily. Adapted from (Kaufmann, 2010).

1.1.2. Transmission

Mycobacterium tuberculosis was discovered and shown to be the causative agent of TB by Robert Koch in 1882. Thirty years later, Canadian physician William Osler noted that "*all who mix with tuberculosis patients got infected, but remained well so long as they took care of themselves and kept the soil in a condition unfavorable for the growth of the seed*" (after Dobbs & Kimmerling, 2008). In the following years the mechanism of airborne infection was partly elucidated by studies on droplet nuclei (Wells, 1934) and the deposition of airborne bacteria in lung was described (Riley & O'Grady, 1961; Wells, 1934). Today, we know that the cascade of TB transmission starts with the source case, which is a TB patient generating infectious air droplets and expelling them by cough, laughter or any other forceful action of respiratory system (Churchyard et al., 2017). The bacteria survive in the air and are inhaled by an exposed individual, who in turn may become infected and can remain latent (LTBI - latent TB infection) or progress to active disease typically within a year after infection or when the organism is challenged, e.g. by undernutrition (Churchyard et al., 2017; Fox, Barry, Britton, & Marks, 2013). Transmission can occur anywhere where an actively infected person meets other individuals and typically takes place in households or working environments affected by TB. Persons in close contact of the patients are particularly exposed to infection and interrupting this transmission way is a crucial step in counteracting the TB epidemic (Churchyard et al., 2017).

1.1.3. Prevention of TB

The most efficient way of reducing transmission as well as morbidity and mortality caused by infectious diseases is vaccination (Kaufmann, Hussey, & Lambert, 2010). So far, the only vaccine used against TB is *Bacillus Calmette-Guérin* (BCG) which is administered to around 80% of infants worldwide. BCG protects against severe forms of disseminated TB yet it fails to provide protection against pulmonary TB, which is the main disease form (Hatherill, 2011). BCG vaccine is based on attenuated *Mycobacterium bovis,* the pathogen responsible for TB in cattle. Currently more than a dozen vaccine candidates against TB are tested on different stages of clinical trials. Two of the vaccines, VPM1002 (Nieuwenhuizen & Kaufmann, 2018) and *Mycobacterium vaccae* (Kaufmann et al., 2010) are being tested for efficacy and are already on the third stage of the clinical trials.

Prevention of TB relies primarily on early detection and vaccination. According to the WHO recommendations (WHO, 2015b), detection of TB requires active screening of individuals at high risk of TB, e.g. HIV positive (HIV+) individuals or close contacts of TB patients. Early diagnosis and treatment efficiently hinder TB transmission. In May 2014, the World Health Assembly passed a resolution called "End TB Strategy" which aims at reducing new TB cases by 80%, TB deaths by 90% and to protect 100% of the families affected by TB from the tremendous treatment costs by 2030 (WHO, 2015a). This goal can only be achieved by dramatic reduction in TB transmission.

1.1.4. Symptoms and diagnosis of TB

Pulmonary TB is the most frequent form of TB. However, in 15-20% of the cases the bacteria invade other sites, causing extrapulmonary forms of the disease. Bacteria can infect the pleurae causing tuberculous pleurisy, scrofula of the neck when they infect the lymphatic system, urogenital TB in the genital and urinary tract, TB meningitis in the central nervous system, or spinal TB (also known as Pott's disease) when the bacteria infect bones and joints (Golden & Vikram, 2005). Extrapulmonary TB occurs more often in the immunosuppressed individuals, children and HIV+ people. The form of TB affecting multiple parts of the body is called miliary or disseminated TB and it consists of up to 20% of extrapulmonary TB cases (Sharma, Mohan, & Sharma, 2016). If not indicated otherwise, in this thesis I focus on pulmonary TB cases.

The main symptoms of TB are: severe cough lasting at least three weeks, chest pain, and blood or sputum presence in the cough. The accompanying symptoms can include weakness or fatigue, weight loss, lack of appetite, chills, fever, and night sweats (Hopewell, 2017). LTBI individuals do not present TB symptoms and do not spread TB.

In 2016, approximately 40% of TB cases were not reported (WHO, 2017). Every year around one third of the global TB burden remains undiagnosed. Low-income countries, where the disease is particularly widespread, still rely on outdated diagnostic technologies which are ineffective and do not detect drug-resistance (Lawn, 2015). Currently, a wide spectrum of diagnostic tools exist and are being developed to become more cost effective, which will help those who are most in need.

Light microscopy of sputum smears remains the most broadly used TB detection method. Every year, close to 90 million individuals undergo a sputum test (Perkins, 2009). This simple, inexpensive method can detect TB rapidly; however at the same time it is not sensitive enough and relies on the exact examination of the sample by a laboratory technician (Lawn, 2015). Only samples containing more than 10,000 bacilli per milliliter (ml) of sputum are recognized as TB positive, therefore the patients with lower bacterial content in sputum (typically HIV co-infected people) remain undiagnosed (Gupta et al., 2013). The specificity of smear microscopy-based TB detection is sufficient in the high TB burden areas but lower in the high-income countries where positive sputum smears are often caused by nontuberculous mycobacterial species. Nowadays the pre-processing of the sputum and fluorescent staining are used to increase the specificity of detection.

The most sensitive way to detect TB is culture-based diagnosis. The processed sputum sample is cultured on the enriched media and the grown cultures are subsequently visualized. However, since Mtb is an extremely slowly growing bacterium, it requires up to 6-8 weeks for a colony to grow sufficiently. Recent developments of this method include the use of selective liquid media and growth indicator systems (Lawn, 2015).

Pathogens are often detected based on antigen presence in body secretions. In the case of Mtb, urine provides a source of Mycobacterial antigens which can be safely analyzed without risk of aerosol formation (Kashino, Pollock, Napolitano, Rodrigues Jr, & Campos-Neto, 2008). The cell wall lipopolysaccharide lipoarabinomannan (LAM), currently the best candidate antigen, can be detected by commercially available enzyme-linked immunosorbent assay (ELISA; Clearview-TB®-ELISA) and more recently as a point-of-care version. Despite the above mentioned advantages of the urinary TB detection, the LAM assay sensitivity is too low for regular clinical implementation (Minion et al., 2011).

If the Mtb bacteria are present in the organism, their DNA can be rapidly detected in the human blood due to specific amplification reaction. Molecular detection of Mtb is possible thanks to the development of a range of methods based on nucleic acid amplification tests (NAATs): polymerase chain reaction (PCR), real-time PCR, isothermal amplification, and strain displacement (Lawn, 2015). NAAT can be accompanied by hybridization methods. These are highly specific, safe for the personnel and fast; moreover, they can also detect drug resistance in the identified bacterial DNA (Lawn, 2015). Their disadvantages include complexity and requirement for sophisticated equipment. The NAAT-based line-probe assays and Xpert MTB/RIF assay have been already endorsed by WHO (WHO, 2008, 2013b). The latter is a compact independent platform for Mtb detection, fully automated and integrated in a user-friendly, easily operated device (Lawn, 2015). Unprocessed clinical samples are purified and concentrated and the real-time PCR is conducted within the automated framework giving the results of Mtb detection and drug resistance within two hours after sample acquisition. The method detects approximately 9 out of 10 cases with a pooled specificity of 99% and is now widely implemented around the world; however the disadvantages include high cost, sophisticated hardware, necessity of computer connection and complicated service (Lawn, 2015). Hence, it is not attainable for poor areas and accessible mostly in laboratory rather than actual clinical settings.

The aforementioned methods detect active TB. Other tests can be used to confirm Mtb infection. Mtb infection is often detected by the so called “tuberculin skin test” (TST) or Mantoux test. A mixture of mycobacterial antigens called purified protein derivative (PPD) which are not species specific is injected into the epidermis and the host previously exposed to Mtb develops a characteristic skin induration within 2-3 days. The reaction diameter is classified into one of the levels: 0-5mm, 5-10mm, 10-15 mm or >15mm (Nayak & Acharjya, 2012). The medical risks of the tested person determine on which level the test result is considered positive. The drawbacks of the TST include relatively frequent false positive results (Starke, 1996). The false positive results are given by the vaccinated people as well as people with nontuberculous mycobacterial infections, and are estimated as 20% of all positive test results (Rabinowitz & Conti, 2010). The false negative results may happen in recently infected patients, immunocompromised patients or malnourished children (Lloyd, 1968). Especially in children, touching and scratching the injected area also causes redness and swelling which can be interpreted as a positive test result. In some cases hypersensitivity to PPD occurs; therefore the

diagnostic centers using the test need to be equipped with epinephrine (Froeschle, Ruben, & Bloh, 2002).

Another molecule measured to detect previous Mtb exposure is interferon-γ (IFN-γ). IFN-γ-release assays (IGRAs) detect the cytokine present in blood after ex-vivo stimulation with Mtb specific antigens: culture filtrate protein-10 (CFP-10) and early secretory antigen-6 (ESAT-6) (Lawn, 2015). Detection of a positive response indicates previous exposure to those antigens by Mtb infection. The results do not differentiate between TB and LTBI, and therefore the method is not used for the standard diagnosis but rather as complementary information (WHO, 2011). Moreover, old age, HIV coinfection and several other characteristics are associated with false negative IGRA results (Nguyen, Teeter, Graves, & Graviss, 2018).

In the advanced disease phase when granulomas have already developed in lungs, they can be detected by X-Ray scan. The detected abnormalities on chest radiographs can be indicative of TB and support the diagnosis, but they do not serve as diagnosis on their own. However, pulmonary form of TB can be excluded by the lung radiography.

All the above mentioned diagnostic tests are characterized by certain advantages and flaws. The ideal test to detect TB would be cost-efficient, rapid, available at the point of care and able to indicate an efficient treatment regime. The characteristics of such still unavailable tests - target products for TB diagnosis - are listed in Table 1.

Table 1. Target products for TB diagnostics
The table has been adapted and modified from the website: "FIND. Because diagnostics matters." (2018).

Problem	Target product
Triage test	
Cough lasting for at least two weeks can indicate active TB disease; however, majority of individuals presenting this symptom do not have TB. A test excluding TB in such patients would reduce the population which needs to undergo further, more expensive testing.	A point-of-care test to exclude TB, which should be a simple, inexpensive and available for first-contact health-care providers to identify those who need further testing.
Point-of-care non-sputum biomarker test	
Sputum smear microscopy is currently used to detect most TB cases, even though it has suboptimal sensitivity and is difficult in case of children and HIV-infected individuals. In the other hand, molecular detection of TB cannot be performed in most microscopy centers.	A rapid, point-of-care, non-sputum-based test detecting all forms of TB by identifying characteristic biomarkers or bio-signatures. The test would be implemented at microscopy centers, easy to perform, robust with minimal sample preparation and operational requirements.
Smear replacement test	
Sputum smear microscopy is currently used to detect most TB cases, even though it has suboptimal sensitivity and is difficult in case of children and HIV-infected individuals. A more sensitive test at the microscopy center level has the potential to improve patient care by (i) reducing transmission by increasing TB diagnosis, linked to treatment and (ii) leveraging existing infrastructure in microscopy centers.	A more sensitive point-of-care sputum-based test to replace smear microscopy for detecting pulmonary TB that is easy to perform and has minimal operational requirements.
Next generation drug-susceptibility test to inform treatment	
Due to the spreading antibiotic resistance TB diagnosis should be supplemented with the indication of efficient treatment regimen.	A rapid drug-susceptibility test that can be used at the microscopy-center level of the health-care system to select regimen-based therapy. Such a novel diagnostic test should ideally include testing for rifampicin, fluoroquinolones, and pyrazinamide and isoniazid resistance.
Test for detection of disease progression	
Diagnosis and treatment of LTBI should be addressed. Around one third of the world population is infected with Mtb. While current diagnostic tests for infection show that an individual has been exposed to Mtb, they poorly predict whether an individual will progress to active TB in the future.	An ideal test of TB disease progression would differentiate patients in the various stages from infection to active TB and may detect the presence or absence of incipient TB.

Tools used to detect TB are imperfect and above all, do not allow prediction of whom of the infected individuals will develop active TB. Currently, significant hopes are being placed on the development of methods based on combinations of host-related markers which would detect or even predict the disease with superior performance (Maertzdorf, Kaufmann, & Weiner, 2014). Such an approach demands analysis of vast amounts of data and application of specialized bioinformatic tools for classification of healthy, infected, and sick individuals as well as for prognosis of risk and treatment outcomes. In the chapter 1.2, I introduce transcriptomic biomarkers as a candidate method for TB diagnostics.

1.2. TRANSCRIPTOME STUDIES IN TB

1.2.1. RNA expression

The genetic information carrying the instructions for growth, development, functions, and reproduction of every living organism is contained in DNA, which is shared across all cells of an organism. The ability of DNA to instruct development of appropriate cells or tissues is mediated through RNA, which is a functional carrier of genetic information. Fragments of DNA are being transcribed into RNA molecules and further instruct protein translation and expression in a manner dependent on the transcriptional regulation – a mechanism that ensures expression of different sets of transcripts according to the tissue, stimuli, and developmental stage (Adams, 2014). For this reason, the RNA expression levels vary between the cells of a particular organism and are responsible for structural and functional differences between tissues even though the DNA of each cell remains the same. The transcription of DNA into RNA is regulated by proteins called transcription factors (TFs), which can activate or suppress a given gene (Adams, 2014). They function through recruiting RNA polymerase to bind to particular gene's promoter region or by blocking this binding. Once bound to the promoter, RNA polymerase enables production of primary RNA transcripts by pairing subsequent RNA bases with complementary DNA bases. In eukaryotic cells the initial transcripts encoding proteins, called mRNA (messenger RNA), are processed and edited after which they ultimately cooperate with a ribosome to produce the expected protein (Figure 2). Transcripts of other types, like tRNA (transfer RNA) and rRNA (ribosomal RNA) convey their functions without involvement of the translation process. The Human Genome Project estimated that the human genome contains about 20-25,000 genes (Human Genome Sequencing Consortium, 2004). Since thousands of transcripts are produced in every cell during every second, there are many mechanisms controlling this process on every stage – starting from initial transcription control, through RNA processing steps up to protein expression and degradation (Adams, 2014). Gene expression is dynamic, which means that the same gene may act differently depending on the circumstances. Therefore, the level of a transcript of a gene can be indicative of a state of the cell and can support information about what is happening with the host – for example, that the host is undergoing an infection or succumbing to a disease. For this reason, several methods of transcript detection and quantification have been developed.

Figure 2 Overview of the gene expression in eukaryotes
Adapted from Leung, Delong, Alipanahi, & Frey, 2016.

1.2.2. Methods of RNA detection and quantification

Depending on the study type, there are two main categories of the RNA quantification methods. The first category encompasses methods directed to measure RNA of predefined transcripts, which is most useful when the investigation is based on a hypothesis involving an already predefined gene (or set of genes). For example, it tests how the level of particular cytokine involved in a disease changes upon infection. The expression of a particular gene can be directly measured using a technique called northern blotting. In this technology the RNA derived from a sample is separated on an agarose gel according to the size, hybridized to a labeled RNA fragment complementary to the gene of interest, exposed and analyzed (He & Green, 2013). mRNA of a particular gene can be also measured by reverse-transcription quantitative PCR (RT-qPCR) (Bachman, 2013). The reverse transcription of an RNA fragment into DNA is followed by quantitative PCR with use of the generated cDNA template and fluorescently labeled nucleotides. The emitted fluorescence is measured, and the initial amount of RNA can be calculated based on the standard curve.

Apart from detecting and quantifying single transcripts it is possible to investigate the transcriptional profile of a cell or tissue. Such profiling can be performed using the mentioned RT-qPCR, tag-based technologies or microarray technology. The tag-based methods include serial analysis of gene expression (SAGE) and RNA-Seq. They are based on quantifying the amount of times with which each short sequence (tag) unique for a transcript is detected in a sample and therefore, provide a relative measure of transcript concentration. RNA-Seq technique generates simultaneously sequence data that can be matched to a reference genome. Additional information that can be gained using this approach is identification of single-nucleotide polymorphisms (SNPs), splice-variants or even novel genes (Stanton, 2001).

In case of microarray, SAGE or RT-qPCR there is no clear-cut rule regarding their categorization - depending on the scientific approach, they can be used both to validate single genes in a hypothesis-driven approach and to screen hundreds of them in a hypothesis generating approach.

The datasets analyzed in this study had been generated using microarray technology, which is a rapid, reliable, and reproducible technology to detect transcript abundance in a high-throughput manner. A microarray is a collection of microscopic spots of DNA fragments attached to a solid surface, e.g. silicon or glass (Simon, Korn, McShane, Wright, & Zhao, 2003). Two available array types, cDNA and oligonucleotide arrays, differ by the type of the immobilized molecules: up to 5,000 base pair (bp) long cDNA molecules in cDNA arrays *versus* (vs) typically 25-mer long oligonucleotides in high-density oligonucleotide arrays (Schulze & Downward, 2001). The DNA fragments can be also imprinted on the arrays in two ways: by spotting of previously synthesized molecules on the glass (spotted microarrays) or by synthesizing oligonucleotide sequences directly onto the array (Simon et

al., 2003). Each spot (called probe) has defined coordinates and contains picomoles of a specific DNA sequence. This sequence corresponds to a single gene and under strictly defined conditions hybridizes with a complementary (target) cDNA fragment derived from the investigated sample and labeled with fluorophore, silver or chemiluminescence(Simon et al., 2003). After the hybridization the intensity of signal emitted by each spot is measured. From this measurement the relative transcript abundance in the target sample is calculated.

The microarray technology can involve single- or double-color arrays (Duggan, Bittner, Chen, Meltzer, & Trent, 1999). In the first case, the microarray is hybridized with cDNA derived from two samples which will be later compared, and labeled with two different fluorophores. The most commonly used dyes are cyanine 3 (Cy3, green), emitting fluorescent signal at 570 nm wavelength, and cyanine 5 (Cy5, red), with a fluorescent emission wavelength of 670 nm. The samples, each labeled with a different dye, are then mixed and hybridized to an array. The signals of each fluorophore are quantified and differentially expressed genes are identified using their ratios (Duggan et al., 1999; Simon et al., 2003).

Preparation of a sample for the hybridization consists of the following steps: extraction of RNA, isolation of mRNA, quality assurance and concentration measurement, reverse transcription to cDNA, amplification and labeling (Macgregor & Squire, 2002). The labeled cDNA is hybridized onto the array under specific conditions defined by the manufacturer in a hybridization oven, washed to eliminate non-specific binding, and then scanned. Subsequently, the image is transformed into a grid where each spot with measurable intensity occupies one field and the pixel intensity of the fields is quantified.

The microarray technology is used for transcriptional profiling in large cohorts, enabling fast, reproducible high-throughput studies. In the last years the technology became cheaper and therefore also more broadly used in low-resource areas, frequently located in the developing countries which at the same time often have direct access to samples from certain diseases. In high-resource laboratories microarray technology has become partially substituted by RNA-Seq which allows transcript identification without prior knowledge and generates more versatile data. Nevertheless, the microarrays still remain a popular technology to investigate gene expression profiles and at the same time present an already standardized and commercialized field. It is nowadays expected to deposit the collected microarray data in one of the databases like Gene Expression Omnibus (GEO), ImmGen database, or ArrayExpress.

Challenges of the analysis of microarray data which have been addressed in this study, include multiple levels of replication in experimental design, statistical treatment of the data, the number of platforms and independent data formats and mapping each probe to the mRNA transcript that it measures.

1.2.3. Whole blood transcriptomic biosignatures

Blood provides an easily accessible source of information about the state of an organism and WB samples remain the primary source of biomarkers of pathology, including infection (Liew, Ma, Tang, Zheng, & Dempsey, 2006). WB cell transcriptome profiles are thought to illustrate a systemic immune response as blood contains cells and molecules of the immune system and is the carrier of metabolites between different tissues (Liew et al., 2006). WB cell composition in mouse and man is not directly comparable given that it varies in the ratio of neutrophils and lymphocytes – neutrophils comprise 50-70% of human and 10-25% of mouse WB cells, while lymphocytes comprise 30-50% of human and 75-90% of mouse WB cells (Mestas & Hughes, 2004). However, states of infection drive changes in blood composition in both types of host such as emergency granulopoiesis and neutrophilia (Berry et al., 2010; Dorhoi et al., 2013; Lowe, Redford, Wilkinson, O'Garra, & Martineau, 2012).

As early as in 1980's, the term 'biomarker' gained popularity in cancer research being used to describe molecules found in serum and potentially useful in the detection of cancerogenic processes (Paone, Waalkes, Baker, & Shaper, 1980). In 2001, the official definition of a biomarker was proposed by Biomarkers Definitions Working Group (Downing, 2000; Paone et al., 1980). According to the definition, a biomarker is "a characteristic that is objectively measured and evaluated as an indicator of a normal biological process, pathogenic process or pharmacologic response to a therapeutic intervention" (Biomarkers Definitions Working Group, 2001). Biomarkers help to identify different diseases and to define the disease or recovery stage of a patient.

Transcriptomic biomarkers can be derived from any model or tissue affected by infection – e.g. from the mouse model of a certain disease or even more narrowly – from macrophages of an infected individual. Those biomarkers found in known and strictly controlled systems (with known infection time point, in a group of inbred mice living under standardized conditions) can be very precise and distinguish between sick and healthy individuals with nearly 100% sensitivity and specificity.

In patients, acquisition of a specific affected tissue or isolation of particular cells is more challenging. Moreover, since biomarkers are meant to help clinical diagnosis they should be derived from a source that can be not only easily and quickly accessed, but also cheaply and efficiently analyzed. Such sources are body fluids and secretions: saliva, urine, and most importantly – blood. Blood accesses all organs and tissues to deliver oxygen and nutrients while collecting end products of cell metabolism and bringing them to the eliminating organs (lungs, kidneys, liver). It is also a carrier of circulating immune cells. The fraction of human blood used for immune system studies is obtained by removing red blood cells by density gradient centrifugation which separates WB into two fractions – above and below the density of 1.077g/ml in the most commonly used Ficoll gradient centrifugation (Miyahira, 2012). The denser fraction is removed, containing erythrocytes and polymorphonuclear cells (PMNs). The remaining part of lower density contains lymphocytes (T cells, B cells and natural killer (NK)

cells), monocytes, and DCs and is referred to as peripheral blood mononuclear cells (PBMCs). In humans, the frequencies of different cell populations of PBMCs vary across individuals with lymphocytes comprising typically between 70 and 90%, monocytes between 10 and 30% and DCs between 1 and 2%. In a healthy organism T cells constitute between 45 and 70% of all lymphocytes, B cells between 5-20%, similarly as NK cells. Further, the T cell compartment is composed of roughly 65% of CD4+ and 35% of CD8+ cells, among which naive and memory cells can be discriminated (Miyahira, 2012).

Investigating WB transcriptomic profiles represents a new approach to diagnostics – looking at host transcriptomic regulation instead of isolating a pathogen or antibodies to identify the exposure. In cases of diseases like TB this capacity can bring significant advances, because the number of bacteria may be undetectable or the infected tissue inaccessible, and the antibody response may not be established at the moment of testing.

1.2.4. Machine learning in biomarker discovery

Considering vast amounts of multidimensional data provided by "-omics" experiments, ML approaches are currently considered the panacea for the identification of patterns in datasets from versatile areas (Jagga & Gupta, 2015). ML, a group of methods from the field of artificial intelligence, provides the potential to mine huge datasets with (supervised ML) or without (unsupervised ML) external information about which samples in a dataset belong to what class of input data (called a "label") (Zhu & Goldberg, 2009). On the example of TB, supervised ML methods are used to analyze the samples labeled as "TB" or "healthy" and learn to classify new, unlabeled samples based on the variables from the training dataset. Unsupervised ML methods find clusters of unlabeled data which may correspond to the classification of donors as "TB" or "healthy". Supervised ML algorithms can be used for classification (predicting a discrete class label output) or regression (predicting a continuous quantity output) problems. They also include a subgroup classified as "semi-supervised" which utilize both labeled and unlabeled datasets (Zhu & Goldberg, 2009). Unsupervised ML algorithms are implemented for clustering, density estimation or dimensionality reduction and class label information if patient sample is not available. The models or classifiers generated by ML can serve as standalone executable systems predicting clinical phenotype of the new patients in clinical decision support (Jagga & Gupta, 2015).

A good ML study should be characterized by careful design which includes non-overlapping training, test datasets, and if possible, an independent validation dataset. An input dataset for ML is a matrix of samples *vs* attributes (features) (Jagga & Gupta, 2015). In case of the work presented here, the samples correspond to blood donors and the features correspond to transcripts measured on microarrays. *Training dataset* is the data subset used to train the learning algorithm to identify *classifier* (in the case of this study: a transcript set). At this stage *cross-validation* is used to estimate the errors

and generalizability of the classifier. There are several statistical procedures which can be used for cross validation: leave one out, hold-out, bootstrapping and k-fold cross validation (Jagga & Gupta, 2015) applied in this study. *Bootstrap* validation is based on subsampling performed with equal replacement from training dataset. In *k-fold cross validation* the dataset is split in k-mutually exclusive subsets, subsequently the classifier is trained on k-1 subsets and tested on the remaining one. The procedure is repeated k times and the average accuracy of k-folds is the estimated accuracy of the classifier. The classifier is then tested using the remaining, untouched data subset - *testing dataset* - thanks to which the performance and error of the classifier can be estimated. The *validation dataset* is derived from another corresponding study and it indicates if the model can be widely implemented. Therefore, additional features of a good ML model are cross comparability and wide implementation. The model should ultimately become publicly accessible for further testing, improvements and wider translation (Jagga & Gupta, 2015).

Building a classifier can be preceded by *feature selection* – a technique used to reduce the dimensionality of the model, improve its performance, avoid overfitting, increase cost-effectiveness and ultimately also to gain insightful clues about the processes described by the data, e.g. disease pathogenesis and progression (Saeys, Inza, & Larranaga, 2007). Feature selection means determining the important features (in the case of this study: biomarkers) which the model should consist of, since the multidimensional "-omics" data is characterized by so called 'curse of dimensionality', which means that almost in every such study there are far more measured features than collected samples (Jagga & Gupta, 2015). For example, microarrays can measure around 40,000 genes while the microarray cohorts normally consist of between 10 and 1000 donors, and far fewer when we investigate an infectious disease like TB. This can lead to *overfitting*, which means that the classification model contains so many irrelevant features that it becomes over sensitive to the investigated training set. Feature selection solves this problem by identifying the meaningful, sensitive and specific disease markers.

Feature selection algorithms can be independent of the classification algorithm, based on the data properties (filter methods), based on the evaluation of the learning models with selected feature subspace (wrapper methods) or estimate the optimal feature subset by grading feature importance within the classification algorithm (embedded methods). The last one has been applied in this thesis and they are less prone to overfitting and less computationally heavy than the wrapper methods while at the same time selected in cooperation with the learning algorithm in contrary to filter methods (Saeys et al., 2007).

A simplified workflow of supervised ML is presented in the Figure 3.

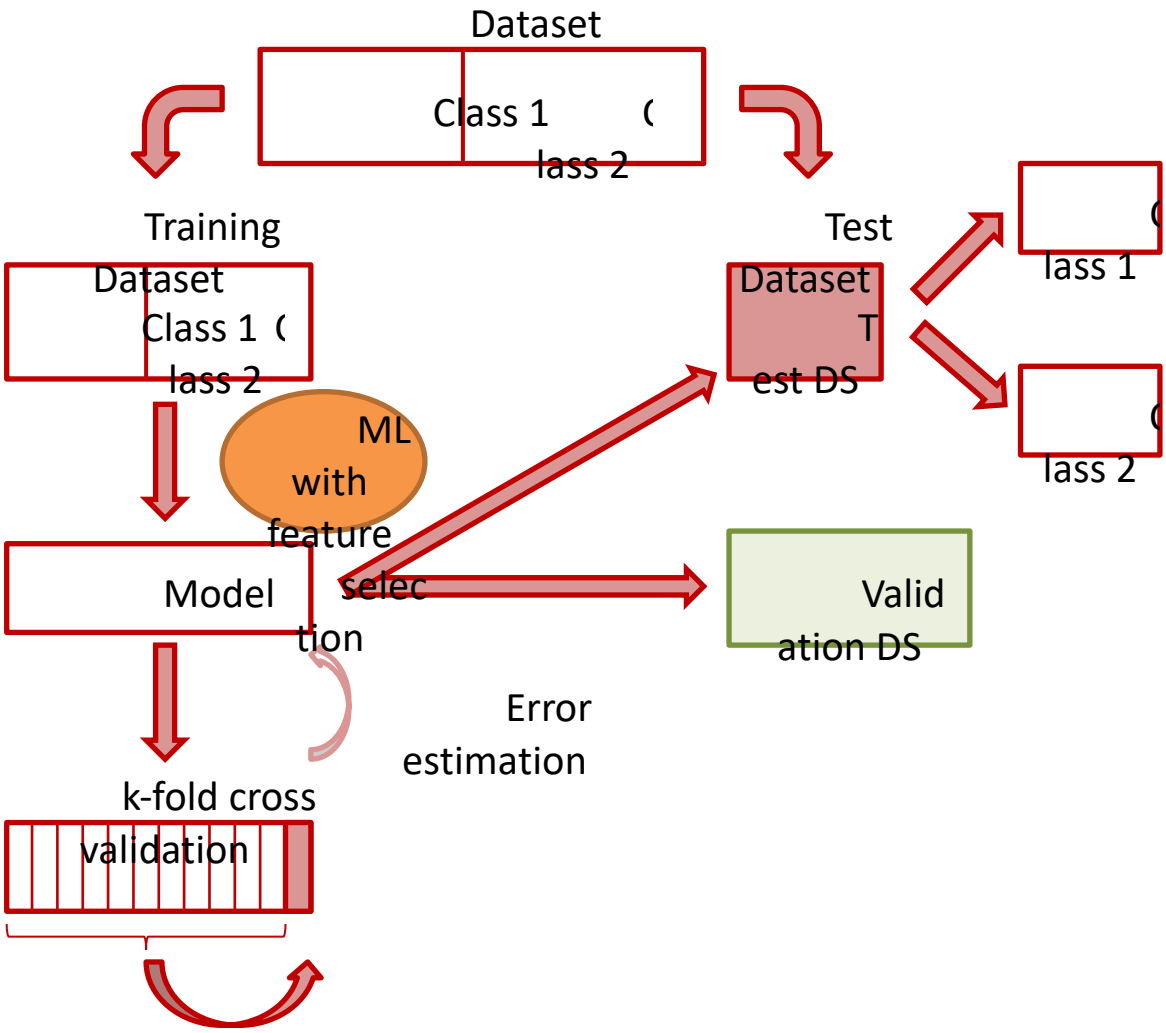


Figure 3 Example of workflow of supervised ML

A dataset with class labels is divided into training and test set. The algorithm learns to classify the samples and uses the embedded feature selection method. The created model is first cross-validated using k-fold cross validation to estimate the errors. Then, the classifier is applied to identify classes in the test and independent validation dataset.

1.2.5. Unsupervised Machine Learning – Principal Component Analysis

Principal component analysis (PCA) is an unsupervised ML method reducing the dimensionality of multi-dimensional data (Shlens, 2014). Reducing the dimensionality simplifies the data and helps to understand it; it can explain where the variance in the data comes from, facilitate data visualization and description.

PCA finds principle components of the dataset, which means the variables, which explain the largest portion of variance in the data. In the case of microarray analysis, it can be used to assess the influence of technical parameters of the experiment on its outcome as well as to find which of the sample characteristics are responsible for variability between patients (e.g. TB status, HIV coinfection or ethnicity as I will present in the Chapter 3). PCA transforms the data into a new coordinate system where the first axis corresponds to the first Principal Component (PC) along which the data presents

greatest variance (Shlens, 2014). Mathematically, the PCs are the eigenvectors of the covariance matrix of the original dataset. Since the covariance matrix is symmetrical, the eigenvectors must be orthogonal to each other. Every eigenvector possesses a corresponding eigenvalue, which is a number indicating how much variance there is in the data along its eigenvector (Shlens, 2014).

Since PCA finds principal components with the greatest variance, data normalization helps to avoid a situation in which the first PC is dominated by variables with large values and large absolute variances and not by the attributes showing major biological differences. Data normalization keeps the variables on similar scale which gives every variable a chance to form part of the PCs (Hamilton, 2014).

1.2.6. Supervised Machine Learning – Random Forest

Random forest (RF) models can be used when the aim of the research is to classify samples into categories (called classes) of interest based on multiple measured or described variables which in biology can include any functional categories like disease status, gender or species. RF can also serve to regress features against quantitative data. They are based on decision trees, which can be used for classification of discrete variables or regression of a continuous variable, therefore being classified as Classification and Regression Trees (CART).

RFs are based on generating many classifiers and aggregating their results (Liaw & Wiener, 2002). Two well-known methods of classification are (i) boosting, where successive trees give extra weight to points incorrectly predicted by earlier predictors (Schapire, Freund, Bartlett, & Lee, 1998)), and (ii) bagging, where each successive tree is created independently of the previous trees using bootstrap of a dataset (Breiman, 1996). In this study I am using a method proposed by Breiman in 2001, in which each tree is constructed using a different bootstrap sample of the data and each node is split using the best among the subset of predictors randomly chosen on that node (Breiman, 2001). This method was shown to perform very well in comparison to many other classifiers and is robust against overfitting. Additionally, it demands only two parameters: the number of variables in the random subset and the number of trees in the forest.

RF algorithm involves randomly subsetting samples from a dataset and building a decision tree based on those samples. At each node in the tree a number of features are selected and the feature providing best split given any preceding nodes is selected. The same principle is applied to multiple trees based on different sample subsets which prevents overfitting. In this study, each tree's performance has been tested on the left-out samples using k-fold cross-validation. Later, the model derived from RF built on the training set with selected features sorted by variable importance was tested on the test set to evaluate the performance of the model and on an independent dataset to assure biological relevance.

RF modeling allows calculating the variable importance of all the features in the model, which creates grounds for biosignature detection. It is performed by re-running the RF algorithm on the same dataset but with one feature value scrambled across all samples and calculating the difference in accuracy between the two models.

A scheme of RF decision making is presented in the Figure 4.

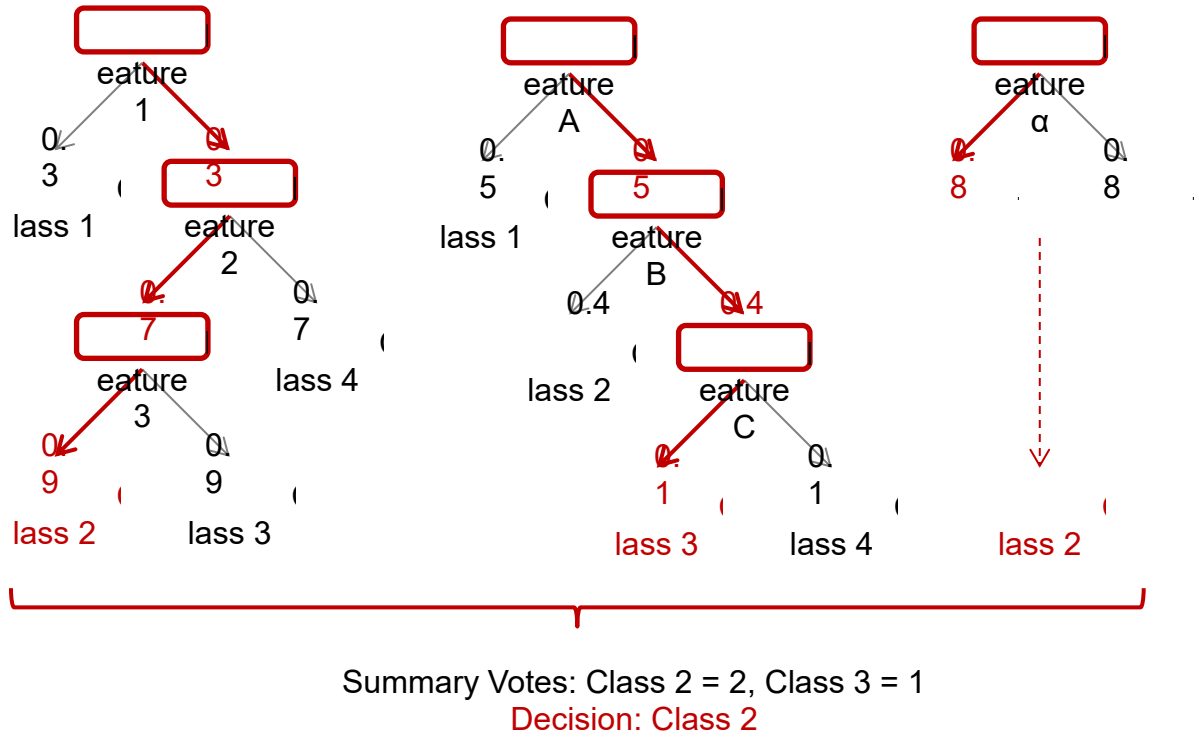


Figure 4 Simplified scheme of the classification RF algorithm

Single decision trees on each node make a decision leading to primary sample classification. Summary votes of the decision trees decide about the final sample classification.

1.2.7. Approaches to identify diagnostic TB biomarkers in published studies

Biomarkers discriminating between TB patients and healthy donors or OD patients have been identified by a number of studies.

As early as 2007, Mistry et al. identified a set of 9 gene transcripts in blood which discriminated between the subjects cured from TB after conventional chemotherapy and the patients susceptible to recurrent disease (Mistry et al., 2007). The signature was based on the WB samples from 10 individuals in each clinical group, all recruited in South Africa (SA). At the same time, Jacobsen et al. compared gene expression profiles of PBMCs from 9 TB patients and 9 LTBI donors recruited in Germany and proposed three other genes: lactoferrin, CD64 and the Ras-associated GTPase 33A for classification of TB patients (Jacobsen et al., 2007). Those first studies on limited numbers of participants were followed by several others involving increasing numbers of patients and different ethnic groups.

In 2010 Berry et al. identified a 393-transcript signature for active TB on patients from the United Kingdom (UK) and SA (altogether 54 TB patients, 105 healthy donors including LTBI) and a specific 86-transcript signature that discriminated between active TB and OD (altogether 193 patients with other pulmonary diseases). The signatures were dominated by IFN inducible genes consisting of both IFN type I and IFN type II signaling (Berry et al., 2010). Using flow cytometry, the study demonstrated that this profile was driven by neutrophils and that at the same time the relative content of B cells and CD4+ and CD8+ T cells was diminished in the TB patients compared to healthy. This change of balance could also influence the blood transcriptional profile differences between the patients and the healthy donors.

In 2011 Maertzdorf et al. recruited 33 TB patients and 34 LTBI donors in SA as well as 9 healthy donors in Germany to define biomarkers predictive of susceptibility and resistance to TB (Maertzdorf, Repsilber, et al., 2011). They identified Fc gamma receptor (FcGR) 1B gene as the most differentially regulated in TB vs LTBI and proposed that together with 4 other genes it forms a signature discriminating the individuals with Mtb infection from the patients with TB disease with high degree accuracy (sensitivity of 94% and specificity of 97%). Among the genes significantly regulated between TB and LTBI determined in the study there was a profound upregulation of Toll-like receptor-associated genes and IFN inducible genes.

In the same year Lesho et al. published a study in which they recruited patients from USA and Brazil who were BCG-vaccinated, latently infected, suffering from TB or healthy and proposed 127 genes capable of accurately classifying samples into the respective four groups (Lesho et al., 2011). They found 13 insulin-sensitive genes to be differentially regulated in all three Mycobacteria stimulated groups which suggested an important role of insulin signaling pathway in TB. Another study by Maertzdorf et al. validated the previously defined signatures on a cohort 46 TB patients, 25 LTBI donors and 37 healthy individuals from the Gambia (Maertzdorf, Ota, et al., 2011). It confirmed the involvement of FcGR 1 signaling in active TB but also pointed out similarities in gene expression profiles of TB and Systemic Lupus Erythematosus patients. Apart from looking at TB biosignature the study approached identification of functional gene clusters playing a role in TB, which resulted in novel insights into the immunoregulatory interactions in TB including the JAK-STAT pathway, microbial sensing by Toll-like receptors and IFN signaling.

In line with these results, in 2012 the whole-genome PBMC expression profiling study by Ottenhoff et al. on 23 TB patients during disease, treatment and after recovery and 23 healthy household contacts from Indonesia emphasized the role of the detected signature of type I IFN signaling in active TB. The authors of the study suggested that the IFN type 1 signaling cascade could be used as a quantitative tool for monitoring active TB. They also showed that the observations acquired from

PBMC profiling were reflected by pulmonary and macrophage response to Mtb infection (Ottenhoff et al., 2012).

A comparison of expression profiles of 8 TB patients, 18 LTBI and 18 sarcoidosis donors in 2012 revealed that the two diseases manifesting in pulmonary pathology with similar histological and clinical symptoms but of different origins share a highly similar gene expression profile (Maertzdorf et al., 2012). Similar to previous studies it included a dominant IFN-inducible gene expression profile. The study pointed out the commonalities and unique signatures in WB gene expression profiles of the two diseases. These observations were confirmed in a study from 2013 (Bloom et al., 2013) which showed that there is a significant difference in the degree of transcriptional activity in TB and sarcoidosis and compared gene expression in those diseases to the transcriptional profiles of pneumonia and lung cancer. This study, conducted on 35 TB, 16 lung cancer, 14 pneumonia and 61 sarcoidosis patients and 113 healthy volunteers identified a set of 114 blood transcripts able to distinguish TB from the three other pulmonary diseases and described the transcriptional response to anti-TB treatment. They as well identified IFN-inducible blood transcriptional signature present in the pulmonary granulomatous diseases, TB and sarcoidosis, as distinct from other lung diseases representing acute and chronic conditions, pneumonia and lung cancer, in which the inflammatory signature was dominant.

Another study dedicated specifically to gene expression changes during anti-TB treatment, was conducted on the WB from 29 treated TB patients and 38 LTBI individuals from SA and 8 treated TB patients from the UK (Bloom et al., 2012). It identified a treatment specific 320-transcript signature which significantly diminished during the first two weeks of treatment and continued to cease until the completion of the 6 months treatment regimen (Bloom et al., 2012). Those finding suggested that blood transcriptional signatures could be used as surrogate biomarkers of successful treatment response. Another study in this direction by Cliff et al. (2013) focused on the networks of genes regulated with time of TB treatment. It underlined relevance of the initial regulation of complement components C1q and C2 followed by slower changes in expression of B-cell markers, transcription factors and signaling molecules. These results were further confirmed by the study of Cai et al. in 2014, where complement gene expression in PBMCs of 9 TB patients, 6 LTBI and 6 healthy controls (HCs) from China was determined using whole genome transcriptional microarrays and presented significant increase in C1q expression in TB patients, which correlated with sputum smear positivity and was reduced after anti-TB treatment (Cai et al., 2014).

In a study from 2013, Kaforou et al. investigated blood transcriptional profiles of 311 South African and 273 Malawian adults including 194 TB cases with and without HIV coinfections, patients with other diseases and HCs (Kaforou et al., 2013). In this study, a 27-transcript signature was proposed to distinguish TB from LTBI and a 44-transcript signature to distinguish TB from OD. Additionally, the authors developed a method for translation of multiple transcript RNA signatures into a Disease

Risk Score (DRS) which can be applied to each patient to evaluate the signatures and developed as a test for TB.

Since TB continues to cause a high toll of disease and death among children worldwide, two transcriptome studies have been dedicated to childhood TB. In the first of them, WB samples were analyzed to identify a minimal set of 9 genes predictive for TB vs LTBI in a group of 9 patients, (Verhagen et al., 2013). Validation of the biosignature on the previously published datasets showed that the biosignature was highly discriminative also for other ethnicities. Functional annotation of the genes suggested a role of calcium signaling and calcium metabolism in active TB. The other childhood TB transcriptomic study was published two years later and involved 114 TB patients, 57 LTBI and 175 OD donors from Kenya and Malawi (Anderson et al., 2014). Once again, the study identified a biosignature of active TB vs OD, which consisted of 51 transcripts; however the authors did not suggest any functional interpretation of this gene set.

Similarly, in 2014 Dawany et al. analyzed global gene expression data from PBMC samples of 43 TB patients from SA with or without HIV coinfection and identified a 251 gene signature that accurately distinguishes HIV/TB coinfected patients from non-TB, HIV-infected patients. The signature diminished as a correlate of the length of anti-TB treatment (Dawany et al., 2014). In 2016 Walter et al. identified a WB TB biosignature among 35 American TB patients in comparison to 35 LTBI and 39 pneumonia donors. However, the accuracy of the classifiers from this study decreased when tested in other populations, which suggested that further investigation is needed to provide generalizability of the signature (Walter et al., 2016). In the same year, Blankley et al. focused his research of finding biosignatures distinguishing TB patients from HCs and from sarcoidosis but with distinction of pulmonary and extra-pulmonary TB (Blankley, Graham, Turner, et al., 2016). They showed that the blood transcriptional responses in pulmonary and extra-pulmonary TB are distinct and reflect the extent of disease symptoms. In yet another approach to find correlate biomarkers of active TB a group from Taiwan investigated microRNA profiles of PBMCs of 7 TB, 7 LTB and 7 healthy donors (Wu et al., 2014) resulting in identification of several microRNA-gene interactions that may serve as potential biomarkers of TB and LTBI.

Biomarkers proved successful not only in identifying TB patients in a population but also in differentiating between infections caused by different Mycobacteria. In a study from 2015 transcriptomic and metabolic profiles of Mtb and *Mycobacterium africanum* (Maf) infected patients were compared to identify host biomarkers associated with lineage-specific pathogenesis and response to anti-TB treatment (Tientcheu et al., 2015). Interestingly, transcriptomic profiles of the 12 Gambian TB patients infected with Mtb and the 14 infected with Maf were similar before treatment – however after treatment over 1600 genes related to immune responses and metabolic diseases were differentially expressed between the two groups. The differences in both PBMC and serum metabolic profiles

between the two clinical groups might be indicative of different treatment efficacy or to individual variability between hosts related to TB susceptibility.

Table 2. List of the TB studies described in the Chapter 1.2.7

The table characterizes the studies with the original publication reference, GEO accession number if available, cohort origin, sample type and the number of donors included in each study (HC – healthy control donors including non-infected, HIV infected and LTBI; Maf – *Mycobacterium africanum* infected donors).

Author	GEO	Cohort	Sample	Cases
(Anderson et al., 2014)	GSE39941	Malawi, Kenya, SA	WB	114 TB, 57 HC, 175 OD
(Berry et al., 2010)	GSE19491	UK, SA	WB	54 TB, 105 HC, 193 OD
(Blankley, Graham, Turner, et al., 2016)	GSE83456	UK	WB	45 TB, 61 HC, 49 OD
(Bloom et al., 2012)	GSE40553	SA, UK	WB	37 treated TB, 38 HC
(Bloom et al., 2013)	GSE42834	UK, France	WB	16 TB, 113 HC, 91 OD
(Cai et al., 2014)	GSE54992	China	PBMC	9TB, 12 HC
(Cliff et al., 2013)	GSE31348	SA	WB	27 TB pre- and post-treatment
(Dawany et al., 2014)	GSE50834	SA	PBMC	21HIV/TB, 22HIV
(Jacobsen et al., 2007)	GSE6112	Germany	PBMC	23 TB, 2 recovered TB, 27 HC
(Kaforou et al., 2013)	GSE37250	SA, Malawi	WB	194 TB, 259 HC, 83 OD
(Lesho et al., 2011)	NA	USA, Brazil	WB	5 TB, 5 BCG, 13 HC
(Maertzdorf, Ota, et al., 2011)	GSE28623	The Gambia	WB	46 TB, 62 HC
(Maertzdorf, Repsilber, et al., 2011)	GSE25534	SA	WB	33 TB, 43 HC
(Maertzdorf et al., 2012)	GSE34608	Germany	WB	8TB, 18 HC, 18 SARC
(Mistry et al., 2007)	NA	SA		10 TB, 10 LTBI, 10 cured, 10 recurrent
(Verhagen et al., 2013)	GSE41055	Venezuela	WB	9 TB, 18 HC
(Ottenhoff et al., 2012)	GSE56153	Indonesia	PBMC	23 TB , 23 HC
(Tientcheu et al., 2015)	GSE62147	The Gambia	WB	12Mtb, 14 Maf pre- and post-treatment
(Walter et al., 2016)	GSE73408	USA	WB	35 TB, 35 HC, 39 OD
(Wu et al., 2014)	GSE62525	Taiwan	PBMC	7 TB, 7HC

1.2.8. Approaches to identify prognostic TB biomarkers in cohorts

Factors associated with the increased risk of progression to active TB include: HIV infection, age, weak immune system, sex, and above all recent contact with a patient with active pulmonary TB. Nevertheless, no diagnostic test to date can indicate if an Mtb-infected person will progress to the disease. Naturally, household, work and medical personnel contacts of a TB patient form a high-risk group and therefore finding a prognostic biomarker of TB to screen those groups will significantly improve the perspectives of decreasing TB burden.

So far, two prognostic biosignatures of TB have been proposed. The first one resulted from following over 6000 adolescents infected with Mtb for two years (Zak et al., 2016). Among the infected, 46 people developed active TB. The suggested biosignature was developed based on those 46 TB patients and 107 matched controls using k-top-scoring pairs ML method and consisted of 16 genes which predicted TB progression with a sensitivity of 66.1% (95% CI = 63.2% – 68.9%) and specificity of 80.6 % (95% CI = 79.2% - 82%) 12 months before diagnosis. The risk signature was tested on the test set and validated on an independent dataset from cohorts from SA and the Gambia from the Grand Challenges 6-74 study including household contacts of adults with sputum-smear positive TB. The study has also shown that it is possible to predict conversion from LTBI to TB up to 18 months before the disease manifestation. Another encouraging conclusion from the study is that the prognostic signature was universal for the investigated patients – since the training and validation cohorts varied in the age and origin of participants.

Another study based on similar assumptions of following healthy household contacts of TB patients identified a four-transcript Pan-African prognostic biosignature of TB (Suliman et al., 2018). Between 3 and 24 months after exposure 79 people progressed to develop active TB in a cohort of almost 5000 HCs. The derived signature predicted progression to active TB up to two years before the diagnosis in the test set consisting of South African, Gambian and Ethiopian study participants. On top of that, the study identified several gene pairs that predict TB progression in various locations in Africa. Out of these C1QC and TRAV27 gene pair consistently predicted TB in adult HCs from multiple sites in Africa (Suliman et al., 2018).

Successful applications of prognostic TB biosignatures could result in targeted treatment to prevent TB. Since around 30% of the population of the world is infected with TB it is not feasible to treat all infected individuals even with the assumption of developing a treatment with minor side effects. Currently used treatments pose risk of serious side effects and are definitely too expensive to be supplied for every infected person as preventive measure.

1.2.9. Approaches to identify universal TB biomarkers in multi-cohort studies

Three studies published in 2016 and 2017 intended to integrate the listed transcriptomic studies. The first study focused on deriving a diagnostic gene set from TB patients which meets the requirements of WHO, including: (i) being derived from non-sputum samples, (ii) maintaining sensitivity of over 80% in patients co-infected with HIV, (iii) maintaining sensitivity of over 66% in children who are TB culture positive, (iv) being simple to conduct (WHO, 2014). The study included altogether 14 WB datasets containing samples from patients from 10 countries, both children and adults. It identified a three-gene set which differentiates between active TB and healthy, LTBI or OD with high sensitivity and specificity (Sweeney, Braviak, Tato, & Khatri, 2016), successfully generalizing the information derived from separate cohort studies. The authors used two meta-analysis methods to perform the study: combining gene expression effect size using DerSimonian-Laird method and combining p-values with Fisher's sum of logs method. They thoroughly investigated the parameters characterizing performance of the three identified genes: KLF2, DUSP3 and GBP5 in diagnosing TB. This study focused on the technical performance of the meta-signature but did not approach interpretation of its biological meaning.

Only four months later another multi-cohort study of TB patients revealed a 380-gene meta-signature of TB vs healthy (Blankley, Graham, Turner, et al., 2016). There, meta-profiling of significantly regulated gene lists was applied to identify the number of overlaps between datasets required for inclusion in the meta-signature. Contrary to the previous study, here modular analysis framework (Chaussabel et al., 2008) was used to identify the common transcriptional responses of patients with TB as biologically meaningful gene modules. This study revealed similarities among all cohorts related to strong upregulation of genes involved in IFN signaling, inflammation, dendritic cells (DCs), apoptosis, cytotoxicity, and under-expression of genes related to B-cells, T-cells, lymphocyte activation and mitochondrial stress (Blankley, Graham, Turner, et al., 2016).

The results were confirmed by a later study in which on the basis of meta-analysis a network of responses to active TB was generated and monitored during treatment (Sambarey et al., 2017). The core of this network consisted of 380 genes among which STAT1, PLSCR1, C1QB, OAS1, GBP2 and PSMB9 were pivotal hubs. Among those hubs, STAT1, PLSCR1, OAS1 and GBP2 are directly involved in IFN signaling pathways. The created network captured biological processes involved in response to TB, including pro-inflammatory responses, apoptosis, complement activation, cytoskeletal rearrangements, cytokine and chemokine signaling.

Insights missing in the multi-cohort studies

In each study, the analysis focused on the trends presented by the TB patients among all the cohorts. It accounted for the variation between different experimental settings and technical differences

providing platform-independent results. However, none of the mentioned meta-studies investigates individual variability of patients belonging to various cohorts. The assumption that the general trends are universally present in the TB patients poses a risk of ignoring alternative events occurring during the immune response to TB in subgroups of patients, for example those with coinfections, weaker immune systems or other unidentified factors. Therefore, in this thesis I investigated the individual variability in responses to TB presented by single individuals trying to answer the question of what types and elements of immune response to TB drive the observed trends.

1.3. VARIOUS FACTORS INFLUENCE MTB INFECTION PROGRESS

1.3.1. Variability in the Mtb infection outcomes

Part of the challenge of combating TB is related to the fact that there are drastic differences between its active and latent form and even further – between the manifestations of active TB in different hosts. TB can be harmful even when characterized by low bacteria number and can be contained even by individuals infected with high bacterial loads.

Even though we know some risk factors of active TB, we do not understand what leads to its development or containment among infected healthy adults. The prognostic biomarkers help to identify the individuals with highest risk of progression (Sweeney et al., 2016) but so far the mechanisms underlying them are obscure. Apart from a binary "TB/LTBI" classification the close look at TB patients reveals huge differences in disease scale, severity and manifestation among patients. As mentioned earlier the most common disease form is pulmonary TB but the bacteria can also invade other organs. There are patients who quickly react to anti-TB treatment and patients prone to relapse. Comparing the disease in humans is of course challenging not only due to the individual variability but also because of hard to control environment and daily routine of the patients. However, even the studies involving macaques living in controlled conditions and infected with equal Mtb doses result in varying disease outcomes (Gideon, Skinner, Baldwin, Flynn, & Lin, 2016). Not only do some of the animals remain latently infected and some progress to active disease, but they also develop disease with varying severity and lung pathology. Other animal models like mice can be divided into strains with low and high susceptibility to TB and in many cases, we do not know what their susceptibility depends on, which will be further discussed in the Chapter 1.4.1.

The patients diagnosed with TB present a spectrum of pathology, ranging from influenza-like symptoms to fully symptomatic disease with blood-stained sputum, weight loss and detectable changes in the lungs. Further investment in analysis of transcriptomic biomarkers for TB early diagnostics is expected to enhance our understanding of susceptibility and resistance to TB (Maertzdorf et al., 2014). For those reasons, host gene expression changes during TB are extensively studied. Discovering the

role and correlates of individual variability in host response to TB is part of the motivation of this thesis and pursued by close investigation of variability of host immune responses among TB patients.

1.3.2. Complexity of the immune response to TB

As soon as an infectious droplet containing sufficient dose of Mtb is inhaled into the airways and alveoli of a susceptible person, the development of Mtb infection begins. The early series of events includes recruitment of phagocytic cells to the site of infection (Schlesinger, 1996). Macrophages and neutrophils contribute to the first-line of defense against TB, but their role in protection is complex and depends on the context of their activation. It has been shown that they can drive both containment as well as progression of the disease thus leading to variable outcomes in the hosts (Lowe et al., 2012; Schlesinger, 1996). Macrophages can contain mycobacterial spread through apoptosis or, opposite, contribute to its dissemination via necrosis followed by infection of the neighboring cells (O'Garra et al., 2013). Neutrophils play a detrimental role participating in granuloma formation and their levels have been shown to be elevated in TB susceptible murine models while their elimination led to increased protection in those animals (Eruslanov et al., 2005; Keller et al., 2006). At the same time, neutrophils promote the development of adaptive immunity against TB by delivering the bacteria to DCs which enhance the initiation of naive CD4+ T cells activation (Blomgran & Ernst, 2011). When the antigen presentation occurs, antigen-specific CD4+ T cell responses are initiated in local lymph-draining lymph nodes, where the T-cell numbers rise and wherefrom the trafficking into the lung begins (Reiley et al., 2008).

The immune response in TB is complex and despite all scientific efforts, so far remains only partially understood which is among others owed to drastically different infection and disease fate depending on the host. Major Histocompatibility Complex class II (MHC II) deficient mice present increased susceptibility to TB which suggests that the CD4+ T-cell mediated immunity is central to TB protection (Caruso et al., 1999). Likewise, patients with HIV infection, characterized by low levels of CD4+ T cells in blood are highly susceptible to TB (Cooper, 2009). CD8+ T cells have been shown to provide protective immunity against TB by production of IFN-γ and enhancing lysis of infected macrophages (Flynn & Chan, 2001). Other T cell subsets, like γδ T cells, CD1-restricted T cells, NK T cells, CD25+ 4+ T cells and Th17 cells play important regulatory roles in the response against TB (Behar & Boom, 2017).

The knowledge about the roles of B-cells in TB containment is also very limited. Since Mtb is an intracellular pathogen, it is expected that the protection is mostly mediated by mechanisms activated in the macrophages by cytokine signaling. Nevertheless, aggregates of B-cell follicles have been observed in the lungs of TB mouse models (Chackerian, Alt, Perera, & Behar, 2002; Maglione & Chan, 2009; Maglione, Xu, & Chan, 2007) and of patients (Ulrichs et al., 2004). Activated B-cells are present in the granulomas of Mtb-infected macaques (Lin, & Flynn, 2012).

Due to the described variability of the roles of particular immune cell types visible already during primary events after Mtb infection I decided to focus my study on deciphering patterns of immune response activation presented by individual patients and by various mouse strains. As the focus and starting point of the analysis of differences presented on transcriptomic level by individual TB patients I chose a signaling pathway considered to play a crucial role in the development and containment of TB: the IFN signaling pathway.

1.3.3. Interferon signaling pathways in TB

IFN-γ and interleukin-12 (IL-12) signaling molecules have been established as central in providing protective T-cell mediated immunity to TB in both human and murine studies (O'Garra et al., 2013), however certain further discussed observations cast doubt on their role. Tumor necrosis factor α (TNFα) and interleukin-1 (IL-1) contribute to the protective immunity against TB and a number of other molecules, identified among others in the transcriptomic biosignatures of TB are being extensively studied (Donovan, Schultz, Duke, & Blumenthal, 2017; O'Garra et al., 2013).

Interferons are proteins released by host cells in response to pathogens or tumors. They are classified into three types based on their cognate receptors: type I IFNs signal through IFN α/β receptor (IFNAR), type II IFN through IFN-γ receptor (IFNGR) and type III IFN through a receptor complex consisting of interleukin-10 receptor beta (IL10R2) and interleukin-28 receptor subunit a (IL28RA) (Platanias, 2005). The three different types of IFN response contribute to TB. The attribution of central role in TB pathogenesis and protection to IFN response has been in recent years supported by the abundant presence of genes involved in IFN signaling in the TB biosignatures identified by the previously listed transcriptomic studies (Berry et al., 2010; Kaforou et al., 2013; Maertzdorf, Ota, et al., 2011).

The type I IFN response has been shown to contribute to either protective or detrimental effects on bacterial infections depending on infectious agent, model system used and acute or chronic state of the infection (Boxx & Cheng, 2016). We do not fully understand the consequences of IFN I signaling in the context of Mtb infection. Data published in several studies suggest that the host resistance and disease severity are influenced by IFN I signaling depending primarily on the host immune competence and possibly - on the pathogen type (Donovan et al., 2017). The TB-susceptible and highly susceptible mice (A129 or 129S2 strains, Il1r$^{-/-}$ mice) survive the infection longer with IFNAR1-deficiency (Dorhoi et al., 2014), which suggests that type I IFN signaling in susceptible hosts infected with Mtb is detrimental. It has been confirmed by other types of studies (Manca et al., 2005; Mayer-Barber et al., 2014; Ordway et al., 2007). In the slowly progressive mouse model for TB, B6D2/F1, administration of IFN α/β antibodies before Mtb infection decreased levels of type I IFNs, increased levels of IL-12 mRNA and decreased STAT1 signaling in the lungs which correlated with longer survival (Manca et al., 2005). This suggests that the reduced levels of type I IFNs lead to upregulation of host Th1

immunity, causing decreased pathogenesis (Manca et al., 2005), which in turn could mean that the IFN response present in the host before Mtb infection influences the infection outcome. In summary, there is a collection of evidence that IFN type I signaling has a detrimental effect on the disease outcome in mice. At the same time, it has been also shown that the IFN-α (subtype of IFN type I) has a protective effect in the absence of IFN-γ signaling (Desvignes, Wolf, & Ernst, 2012; Moreira-Teixeira et al., 2016). To make it even more complex, there are several clinical reports where administration of antimycobacterial antibiotics together with inhalation or subcutaneous administration of IFN-α improved clinical outcomes of the patients (Bax et al., 2013; Giosue et al., 1998; Mansoori, Tavana, Mirsaeidi, Yazdanpanah, & Sohrabpour, 2002; Palmero et al., 1999; Ward et al., 2007; Zarogoulidis et al., 2012). In the patients with genetic lack of IFN-γ signaling, administration of IFN-α improved also the clinical outcome of the patients with mycobacterial non-tuberculous infections (Bax et al., 2013; Moreira-Teixeira et al., 2016; Ward et al., 2007).

IFN type II response pathway involves IFN-γ. This molecule stimulates adaptive immune responses critical for the defense against intracellular pathogens such as Mtb (Bach, Aguet, & Schreiber, 1997). The primary source of IFN-γ are CD4+ and CD8+ cells and recently it has been shown that the innate lymphoid cells, γδ T cells, NK T cells and NK cells can also produce IFN-γ in response to Mtb infection (Bach et al., 1997; Elemam, Hannawi, & Maghazachi, 2017). The secondary producers of IFN-γ play an important protective role in humans (Skeen & Ziegler, 2018).

IFN-γ promotes cellular proliferation, cell adhesion and apoptosis, activates the NK cells, increases antigen presentation and lysosomal activity of macrophages and activates nitric oxide synthase (iNOS) (Zuñiga et al., 2012). At the same time, it can also transfer anti-inflammatory signals to limit inflammation in neutrophils in the chronic infection phase by inhibiting IL-17 production (Zuñiga et al., 2012).

An important contribution of IFN-γ in TB is related to its activity as a mediator of macrophage activation. The dogma of its central role in protection against TB is based on the studies where mice with IFN-γ gene targeted disruption developed granulomas, but failed to produce reactive nitrogen intermediates and restrict bacterial growth (Cooper et al., 1993; Flynn et al., 1993). The mice with IFN-γ knock-out presented heightened tissue necrosis and rapidly succumbed to TB. Administration of recombinant IFN-γ delayed but did not reverse this process (Flynn et al., 1993). Later it had been shown, that in the infected macrophages, IFN-γ induces g-reactive oxygen species (ROS) and reactive nitrogen species (RNS) formation, thus enhancing containment of bacterial proliferation and regulation of intracellular signaling (Cooper, Adams, Dalton, Appelberg, & Ehlers, 2002; Nathan & Shiloh, 2000).

It is still unclear if the role of innate lymphoid cells, γδ T cells, NK T cells and NK cells contribute to the control of Mtb infection through IFN-γ production when adaptive immune response already occurred. In 2006 it was demonstrated that Mtb stimulates NK cell-dependent IFN-γ production

in naive splenic cultures and in lungs of infected mice (Feng et al., 2006). In T cell deficient $Rag^{-/-}$ mice NK cells producing IFN-γ were responsible for the partial resistance of the animals to Mtb infection; however, depletion of NK cells in T-cell sufficient wild-type mice did not influence the infection development.

A number of transcriptomic studies emphasize the regulation of IFN related genes as hallmark of TB. In the study by Berry et al. (Berry et al., 2010) a neutrophil-driven IFN-inducible genetic profile was used to distinguish between TB patients and other patients as well as healthy individuals. In this case both IFN type I and IFN type II inducible genes were included in the biosignature. In 2011, a study on PBMCs investigating differential expression of genes between TB and LTBI showed IFN-γ signaling pathways as significantly differentially regulated (Lu et al., 2011). A study by Harari et al. (2011) investigated the profile of cytokines: IFN-γ, TNF-α and IL-2 in Mtb specific CD4+ T cells in patients with TB and LTBI concluding that the proportion of single-positive TNF-α specific CD4+ T cells was increased among active TB patients, making this parameter the strongest predictor of diagnosis of active *vs* latent TB (Harari et al., 2011). Interestingly, in the recent study it has been shown that the growth arrest of Mtb can be achieved by T cells without IFN-γ production. During the first 21 days after aerosol infection, the growth arrest of Mtb occurred even in the animals unable to secrete IFN-γ (Gallegos et al., 2011). There is evidence that in vitro generated memory CD4+ cells can produce innate cytokines and chemokines providing protection after exposure to an antigen, independent of IFN-γ and TNF-α stimulation (Strutt et al., 2010), however we do not know if this mechanism plays a role in response to TB.

One of the hypotheses explaining the inability of a host to eliminate Mtb is suboptimal production of IFN-γ in the lungs of infected animals. Bold et al. and Winslow et al. (Bold, Banaei, Wolf, & Ernst, 2011; Winslow, Roberts, Blackman, & Woodland, 2003) postulated, that even at the peak of immune response the frequency of activated CD4+ cells secreting IFN-γ is low and decreases during the chronic phase of the infection which favors Mtb persistence. Frequency of IFN-γ producing cells correlates with the availability of the antigen (Bold et al., 2011). In granuloma, the cells undergoing migration arrest can produce cytokines to closely adjacent infected cells, which suggests also a role for T cell derived IFN-γ in activation of infected phagocytes (Bold et al., 2011).

IFN type I and IFN type II pathways often cooperate to efficiently activate innate and adaptive mechanisms leading to immune response (Pestka, Krause, & Walter, 2004). Observations made in IFN-γ signaling deficient patients and $IFNGR^{-/-}$/$IFNAR^{-/-}$ mice encourage further investigation of the mechanisms of this interaction in the context of TB (Desvignes et al., 2012; Moreira-Teixeira et al., 2016). For these reasons, I decided to give attention to those two signaling pathways when investigating individual variability among TB patients.

1.4. THE ROLE OF MOUSE MODEL IN UNDERSTANDING HUMAN IMMUNE RESPONSE IN TB

This part of the introduction has been adapted from my publication published in September 2017 in the Scientific Reports (Domaszewska et al., 2017).

The understanding of pathophysiology of the infectious diseases has been largely broadened thanks to the use of animal models. Soon after the discovery of Mtb as infectious agent causing TB in the 19th century Robert Koch conducted first animal experiments to investigate TB transmission by injecting cultures of Mtb into mice (Gupta & Katoch, 2005). In the following years rabbits, guinea pigs and rats were shown to be susceptible to TB and soon it became apparent, that the reaction to Mtb infection is species-specific. Today, the most frequently used animal models of TB include mouse, zebrafish, guinea pig, rabbit, and non-human primates.

1.4.1. Mouse models of TB

Mouse, the animal of choice for immunological studies for the last century, has markedly broadened our knowledge of the structure and function of the mammalian immune system and understanding of disease mechanisms. Mice share mammalian organ systems present in man, their genome is well described, they breed relatively fast without excessive maintenance costs. Even though mouse and man are divided by 65 million years of evolutionary distance and the evolutionary pressure on immune system is high (Cagliani & Sironi, 2013), it is remarkable how the principles of the immune systems of these two species remain similar (Mestas & Hughes, 2004).

The main discrepancies between murine and human immune systems include the different composition of blood, which is a carrier of immune system molecules. While human blood is rich in neutrophils, which consist 50-70% of all blood cells, mouse blood presents a predominance of lymphocytes, which content reaches up to 90% of total blood cells (Mestas & Hughes, 2004). Another important difference is the repertoire of immune signaling molecules and receptors in the two organisms and broader repertoire of T and B cells in man. For example, the antimicrobial peptides called defensins in humans are produced mainly by neutrophils, while murine neutrophils do not express defensins (Risso, 2000). In contrary, while murine Paneth cells of small intestine express over 20 different defensins, human guts can only produce two types of them. Another example is the molecule CD89, which is an important IgA receptor expressed by human neutrophils, eosinophils, macrophages, DCs and Kupffer cells but absent in mice, which likely use another receptors to bind IgA (Monteiro & van de Winkel, 2003).

The key players in both innate and acquired immune response to TB are mononuclear phagocytes (Dorhoi & Kaufmann, 2015). In a healthy person, the number of macrophage precursors in

blood equals on average 200,000/ml of blood, i.e. 5–6% of the total white cell count (Krikorian, Marshall, Simmons, & Stratton, 1975), whereas in mice around 60,000/ml of blood, constituting around 6% of circulating leukocytes (Krikorian et al., 1975; Sunderkotter et al., 2004). When the bacteria invade the lungs, tissue resident macrophages engulf and constrain them while in parallel recruiting the circulating monocytes and other leukocytes to the site of infection. This contributes to an influx of immune system cells to the lung and ultimately leads to the formation of granulomas in humans and granuloma-like lesions in mice. Therefore, mononuclear phagocytes are one of the most important cell types involved in TB pathogenesis and protection.

Thanks to the mentioned similarities and despite all listed differences, the mouse provides a valuable model to study the response of mammalian immune system to Mtb infection. Owing to the good annotation of the mouse genome, both forward and reverse genetic approaches can be used to define the role of particular genes in the development of TB (Cooper, 2014). In the forward genetic screening approach, random mutations are generated in mice and their influence on TB is investigated. A broad repertoire of mouse strains used to mimic human TB and presenting varying susceptibility can also serve to compare roles of genetic traits in the susceptibility and disease development. Currently, the most widely used mouse strains susceptible to TB include the CBA, C3HeB/FeJ, also called Kramnik's mice (Kramnik, Demant, & Bloom, 1998; Kramnik, Dietrich, Demant, & Bloom, 2000), DBA/2, I/St and 129SvJ, whereas the strains like C57BL/6, A/Sn and BALB/c are resistant (Driver et al., 2012; Medina & North, 2001). Crossing the strains resulted in the identification of several loci responsible for susceptibility, e.g. sst1 containing Ipr1 gene and loci Trl1-4 (Pan et al., 2005; Sánchez et al., 2003); however, as mentioned earlier, the mechanisms of susceptibility are not fully understood yet. I describe the susceptible 129S2 and the resistant C57BL/6 mouse strains in more detail further in this chapter.

1.4.2. Mouse models have advanced the understanding of human TB

Since it is possible to infect mice by aerosol which mimics human infection, we can study the early events after Mtb infection. In mice, the role of particular cell subsets in those early time points are studied by targeted depletion of cells (Kühn & Torres, 2002; Saito et al., 2001).

Understanding the mechanisms of recognition of Mtb antigens and antigen-specific T-cells expansion is furthermore crucial for vaccine design, and here also the mouse model plays a pivotal role (Cooper, 2014). Since the early events following the aerosol infection are not well described, presence of the antigen can only be measured indirectly (Cooper, 2014). This is another area of research where the mouse model broadens our understanding of TB by enabling investigation of the T cell function in the site of infection. For example, use of mouse model confirmed the pivotal role of TNF-α-derived signals in T-cell recruitment and granuloma structure maintenance through its effects on uninfected macrophages (Egen et al., 2008). In another mouse model study it became clear that only small fraction

of Mtb-specific T-cells undergo migration arrest in granulomas, which results in a limited production of the immune signaling molecules (Egen et al., 2011). A study by Bold et al. correlated low level of antigens in granuloma with decreased cytokine production (Bold et al., 2011; Egen et al., 2011). Altogether, these observations led to the conclusion that T-cell location in relation to Mtb-infected macrophages is important for cytokine-mediated infection control.

Another advantage of the mouse as an animal model for TB are established bone marrow chimera models in which bone marrow from the hosts differing in ability to express particular molecules is transferred to an irradiated host and repopulates the hematopoietic system (Cooper, 2014). Such a model has been used to show that within the same host, only the Mtb infected macrophages expressing MHC class II molecules managed to decrease the bacterial burden in a T-cell-dependent manner (Srivastava & Ernst, 2013).

Even though different mouse models are characterized by different lung pathology upon Mtb infection, the human-like mouse granulomas also helped to understand the human pathology. For example, an observation of B-cell follicles associated with inflammation caused by Mtb infection in mice helped to broaden the knowledge of B-cell role in Mtb infection, which includes IL-17 regulation-based reduction in neutrophil influx to the infected lung (Kondratieva et al., 2010; Kozakiewicz et al., 2013; Srivastava & Ernst, 2013). Studying the role of T-cells in granuloma development in the mouse models showed, that despite T-cells low ability to undergo migration arrest, those cells can increase the inflammation in the infection site through macrophage activation. The early observations suggested that the molecules responsible for Mtb control are IFN-γ and IL-12p40 (Cooper et al., 1993; Cooper, Magram, Ferrante, & Orme, 1997), however those observations have been later questioned since in humans the increase in the IFN-γ producing T-cells did not correlate with improved infection outcome. Using gene-deleted and bone marrow chimera mice later on proved that T-cell produced IFN-γ prevents the accumulation of neutrophils and regulates IL-17 activity in the inflamed site, as well as it prevents the accumulation of activated T-cells in the lesions (Desvignes & Ernst, 2009; Nandi & Behar, 2011; Pearl, Saunders, Ehlers, Orme, & Cooper, 2001).

Apart from all of the advantages of the mouse as a small and sustainable in-vivo model of TB, an additional opportunity is its use to test drug and vaccine efficacy and action (Cooper, 2014). In this case, due to the broad repertoire of available strains and mutants, it is up to the investigator to choose the most suitable model. The challenge of the study design lies in the choice of mouse in which crucial features influenced by the tested drug or vaccine precisely reproduce the human system.

1.4.3. Murine models of TB: 129S2 and C57BL/6

Murine models of TB include a broadly used low susceptible C57BL/6 strain and the highly susceptible 129S2 strain (Medina & North, 2001) on which I focus in this thesis. Even though the precise mechanisms responsible for the varying phenotype of those two mouse strains after infection

with intracellular pathogens have not been elucidated yet, it is speculated that the occurrence and the scale of IFN triggered inflammation as well as several identified genetic differences are among the contributors (Davidson et al., 2014; Dorhoi et al., 2014; Govoni et al., 1996; Howes et al., 2016; Kayagaki et al., 2011). The susceptible 129S2 mice succumb to TB within 40 days post infection (p.i.) which is related to excessive IFN type I signaling (Dorhoi et al., 2014), whereas the C57BL/6 mice remain healthy and control the infection up to 300 days (Dorhoi et al., 2014; Medina & North, 2001; Turner et al., 2001). While the lesions formed by the C57BL/6 strain are small and organized, with necrosis observed only at the very advanced infection stage, the 129S2 strain develops necrotic, human granuloma resembling structures (Beamer & Turner, 2005). One of the aims of this thesis is to assess, which of the human immune responses to TB are mimicked by the described mouse strains and what are the reasons underlying their different susceptibility to TB.

1.4.4. Challenges related to the use of animal models

With the rise of high throughput genetic technologies the accuracy of mouse model has been massively questioned (Lin et al., 2014; Mestas & Hughes, 2004; Seok et al., 2013; Shay et al., 2013; Takao & Miyakawa, 2014). On the transcriptomic level the two organisms reveal differences in some of the immune response elements, e.g. cytokine level or NK-cell signaling which, however, is difficult to directly associate with a phenotype. The comparison between heterologous data derived from species-specific experimental settings and different technology platforms, as is the case for human and murine studies, remains a challenge because the data cannot be aggregated and evaluated within a simple statistical framework. This can lead to controversial findings, as described in the following.

In 2013 a comparison of transcriptional profiles of seven non-stimulated murine and human cell lineages collected during immune system development showed similar global expression profiles of corresponding cell types in mouse and man (Shay et al., 2013). A year later another study presented differences in the transcriptional landscapes of the two organisms by describing groups of genes which were tissue-specific or ubiquitous, and identifying a subset of the latter driving the species-specific expression (Lin et al., 2014; Shay et al., 2013).

In 2013 and 2014, the first studies comparing murine and human response to immune system stimulation on the transcriptional level resulted in contradictory verdicts about similarity of gene expression regulation in the two species (Seok et al., 2013; Takao & Miyakawa, 2014). Both studies were conducted using the same datasets from total blood leukocytes from patients and corresponding murine models and applied a correlation approach to identify the similarities. However, not only the biological but also statistical assumptions of the two groups were incompatible, leading to markedly different findings. The first group assumed that the comparison of murine and human gene expression should be performed using all mice and measured features, while the second group hypothesized that due to evolutionary differences murine models should mimic human disease only partially and therefore

chose the most appropriate mouse model and selected only the genes, which were differentially expressed in both species. Seok et al. (2013) implemented Pearson's correlation coefficient in the comparison, while the comparison by Takao et al. (2014) was based on the Spearman's correlation. Recently, another study approached the identification of corresponding immune responses in mouse and man by collecting over 5,000 immune system-specific co-regulated gene sets based on publicly available datasets from mice and men (Godec et al., 2016). This collection defines gene modules regulated concordantly in immunologically relevant comparisons of various cell-state perturbations and diseases from either human or murine studies and suggests how to identify the genes which drive phenotypic differences in both species.

So far, a universal method to compare transcriptome profiles from heterogeneous datasets has been missing. The existing approaches interpret lack of evidence for similarity as evidence of lack thereof and aim at detecting concordances disregarding possibly existing discordantly regulated elements of the immune response of two organisms. The vast collection of co-regulated genes applies exclusively to human and murine studies (Godec et al., 2016). The correlation coefficients-based methods create a risk of identifying a group of discordantly regulated genes as similar if they show a positive correlation coefficient. In analogy, relying on direction of regulation (up- or down-regulated) alone to define similarity of gene expression disregards the precision of the estimated changes in gene expression, including confidence intervals, p-value and effect size which are the indicators of biological importance of the genes regulated in a particular disease.

In this thesis, I introduce a method which allows identifying highly concordantly as well as highly discordantly regulated gene sets between two organisms. The method is based on measuring concordance using directionality of change weighted by the magnitude of gene expression change in two heterologous datasets (for example, human and murine) and associated precision of its estimate. To this end, the approach combines a novel measure of similarity with GSEA and is validated by a simulation study as well as by identification of known similarities between datasets.

1.5. GENE SET ENRICHMENT ANALYSIS REVEALS THE BIOLOGY BEHIND TRANSCRIPTOMIC PROFILES

This part of the introduction has been adapted from the publication co-authored by me, published in September 2016 in the PeerJ Preprints (Weiner & Domaszewska, 2016).

Functionally related, co-regulated or interacting genes are collected and annotated by biologists in clusters called gene sets. Gene sets derived from already described studies, related to known biological effects enable prediction of programs activated by an organism in novel datasets. In the context of gene expression studies in infectious diseases, transcriptomic profiles discriminating between infected and healthy individuals can help to monitor disease progression and reveal or compare

immune responses against pathogens. For this reason, I applied GSEA broadly in my research, to understand the biological background of the structure of the analyzed datasets as well as the analysis outcomes.

Several studies attempted to determine meaningful relationships between genes based on their co-expression under various environmental conditions (Bar-Joseph et al., 2003; Liu, Jessen, Sivaganesan, Aronow, & Medvedovic, 2007). An important strategy for blood microarray data analysis by identifying genes co-expressed across multiple disease conditions and classifying them into 28 blood transcriptional modules was developed in 2008 (Chaussabel et al., 2008). Later, 334 BTMs were annotated according to biological function or tissue-specific expression (Li et al., 2014). Those two sets of BTMs have proved to successfully identify immune responses e.g. for autoimmune diseases or to pyogenic bacteria in patients carrying mutations in pathways responsible for pathogen sensing (Alsina et al., 2014; Pascual, Chaussabel, & Banchereau, 2010). Other widely used gene collections used in transcriptomic studies are gene ontology (GO) sets (Ashburner et al., 2000; The Gene Ontology Consortium, 2017), signaling pathway related sets such as KEGG pathways (Minoru Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2017) and the collections available in the Molecular Signatures Database (MSigDB; Subramanian et al., 2005). Given a wide scope of biological annotation and manual curation, the custom-created gene sets can serve to functionally annotate observed changes in gene regulation, pre-select transcripts crucial for a disease development and immune response and they can be used as a diagnostic tool for detection of disease (Berry et al., 2010), its progression, possible treatment outcomes or best vaccination type.

A typical preliminary output of transcriptomic dataset analysis in a list of genes with their associated fold-changes and p-values in comparison between two conditions (e.g. disease and healthy). This list can serve as an input to GSEA which can be performed using the above-mentioned gene sets (in case of this thesis - mostly BTMs, if not described otherwise). In a first commonly used approach, the list of genes is divided into two groups: “foreground” with genes differentially regulated between the investigated conditions and “background” containing the remaining genes. Then a hypergeometric test is used to test for enrichment with the null hypothesis that there are no more genes belonging to a given module among the foreground than among the background genes. The drawback of this approach is a necessity of setting an arbitrary p-value or fold change threshold to define the foreground and the background, which influences the GSEA results and which results in incomparable results between studies with different sample sizes.

A similar problem occurs if the GSEA is calculated based on statistics from differential analysis, for example by combining the p-values obtained for genes using Fisher’s method. In another approach the initial list of genes is ranked and ordered by the changes between experimental conditions and enrichment occurring towards the top of the list is detected, like in the widely-used GSEA analysis

of MSigDB collections using randomization tests (Subramanian et al., 2005). Here, sufficiently large sample size and high memory and CPU are required for efficient module detection due to use of randomization. GSEA is also not easily compatible with differential expression analysis, for example from the R package limma (Ritchie et al., 2015).

Due to pitfalls of the above mentioned methods in this thesis, I implemented GSEA approach integrated in a novel R-package *tmod* (Weiner & Domaszewska, 2016) which endorses statistical test for enrichment on the ordered lists of genes and is based on an analytical solution rather than permutation or randomization. Tmod testing is suitable for integration with multivariate approaches and it supports use of the BTMs as well as any other custom-created gene sets. Importantly for the presented meta-analytical study as the one presented in my thesis *tmod* contains visualization strategies allowing comparisons of GSEA results among different time points and conditions. The package provides a choice of statistical methods to assign the significance of enrichment analysis including the hypergeometric test, Mann-Whitney "U" test and CERNO test (Yamaguchi et al., 2008). Both U test and CERNO test are performed on the ranked gene lists. The null hypothesis of the U test is that the genes belonging to different gene sets are distributed equally along the gene list and therefore that the mean ranks of the genes from every module are comparable. *tmodCERNOtest* originates from Fisher's combined probability test. It gives higher importance to lower ranking genes which results in p-values better corresponding to the observed effect size. In effect, modules with small effect but containing more genes get higher p-values than in case of the U-test. To overcome the biases of the hypergeometric test, in my analyses I used the highly sensitive CERNO test.

1.6. MOTIVATION

This thesis consists of two parts, both of which are related to analysis of host transcriptomic data in TB and are influenced by the variety of symptoms and outcomes caused in hosts by Mtb infection.

In the first part, I collected publicly available datasets of TB patients and analyzed them together in a single meta-analysis framework designed to identify individual variability in the response to TB presented by TB patients with use of GSEA, PCA, correlation networks and RF ML approaches. The aim was to investigate whether (i) the common trends described in the single transcriptomic studies of TB patient cohorts as well as in the meta-analyses capture variants of transcriptomic programs activated upon encounter with Mtb by individuals or whether (ii) they are representative only for the elements of immune response regulated in the highest order of magnitude, ignoring the immune response elements which are modified more subtly but at the same time influence the disease outcome. The answer to this question can change our current understanding of TB since presence of different types of immune response to Mtb infection in different hosts would mean that not only the scale of pathology can differ in individuals, but also the disease landscape, categorizing it as either inflammatory disease, immunodeficient or rather autoimmune reaction depending on the interaction with the individual host.

In the second part of this thesis, I collected human and murine datasets from TB patients and healthy donors acquired from publicly accessible sources and experimentally by my colleagues from the MPIIB, Department of Immunology. The murine datasets were derived from two different frequently used mouse models of TB presenting drastically different outcome of Mtb infection- one being susceptible and the other one resistant to TB. I used the acquired datasets to answer the question whether the used mouse strains are appropriate models of human TB. I assigned orthologous gene pairs between human and murine datasets and tested the previously described correlation methods to delineate similarity in gene expression between man and mouse but did not obtain any significant results. Therefore, I aimed to establish a method of comparison of human and murine datasets in order to answer the question as to which of the used murine models of TB closely mimics the active disease in man on transcriptomic level. In particular, I wanted to investigate if the similarity between human disease and its animal models can be assessed in a binary way- “similar” or “dissimilar” or if the evolution-based difference between the organisms implies that there are elements of immune response which are regulated in similar way and others which present independent or even discordant regulation. With a robust statistical method this question could be addressed in the context of the investigated mouse strains and human data as well as applied to other dataset comparisons.

2. CHAPTER 2: METHODOLOGY

The challenge in the analysis of datasets derived from various cohorts or different organisms, investigated with diverse study designs, and using distinct technologies lies in the data integration. It should allow the analysis of different datasets in a coherent, single pipeline and at the same time preserve the meaningful information contained in every separate study. Selection of datasets included in meta-analysis must assure data compatibility. It causes difficulties if the expression measurements have been conducted on different microarray platforms, for example when the arrays measured the expression of non-overlapping genes. Comparison of the gene expression data from different species adds another level of complexity imposed by genome evolution involving gene deletion and duplication events. This results in a challenging task of mapping the genes which cannot be straightforward mapped one-to-one. The aim of this section is to introduce and document the methods which I used to integrate the human datasets derived from different studies and to introduce the disco.score – a method created to compare gene expression data from different species. I list and briefly characterize the acquired datasets and explain the implemented normalization procedures. Next, I describe exploratory data analysis which was performed to understand the influence of population-related and technical factors on the differences seen between immune responses of TB patients. I explain the used unsupervised and supervised ML techniques to further explore the biological consequences of the observed variability among the TB patients. Last, I introduce and describe the validation of the disco.score.

2.1. OVERVIEW

In this chapter I describe the implemented data analysis approaches which led to the identification of individual variability in immune response to TB among human patients and the methodology which allowed the identification of concordant and discordant elements of immune response to TB in mouse and man.

All the subsequent data analysis steps have been performed in R programming language for statistical computing (R Core Team, 2018) and the code is available upon request. All the included datasets are publicly available on GEO database (Edgar, Domrachev, & Lash, 2002). The subchapters: 2.2, 2.3 and 2.4 describe the data acquisition, preprocessing and normalization steps common for both projects. The subchapters 2.5 to 2.16 describe the analysis of individual variability among TB patients. The subchapters 2.17 to 2.20 describe the methods implemented in comparing murine with human transcriptomic datasets.

Datasets collected within BioVacSafe project were kindly shared with me by Jeroen Maertzdorf and January Weiner.

The experimental procedures described in the chapters 2.2.4, 2.2.5, 2.2.6, 2.2.7, and 2.2.8 have been performed by my colleagues Lisa Scheuermann, Anca Dorhoi, Karin Hahnke from MPIIB, Department of Immunology and Hans Mollenkopf from Microarray Core Facility of MPIIB, and described by Lisa Scheuermann and Anca Dorhoi. Those listed sub-chapters are identical as in the publication which we published in September 2017 in Scientific Reports (Domaszewska et al., 2017).

The datasets generated during the current study are available in the GEO repository under accession ID GSE89392. The created R-package disco is available on CRAN, under the link: http://cran.r-project.org/web/packages/disco/. The created data collections and module sets can be accessed on the website: http://bioinfo.mpiib-berlin.mpg.de/TBprofiles/.

2.2. DATA ACQUISITION

All datasets used for multi-cohort analysis of human WB transcriptomic profiles of patients with TB are publicly available in GEO data repository (Edgar et al., 2002). The datasets have been acquired using the R-package *GEOquery* (Davis & Meltzer, 2007). Out of the datasets used for comparison of human and mouse transcriptomic responses to TB five were publicly available on GEO. One dataset was acquired experimentally and kindly shared with me by Anca Dorhoi from MPIIB, Department Immunology, and one dataset was acquired experimentally according to jointly planned experiment by my colleague Lisa Scheuermann from MPIIB, Department of Immunology. The microarray sample preparation was performed by Karin Hahnke from MPIIB, Department of Immunology and microarray experiments by Hans Mollenkopf in Microarray Core Facility of MPIIB

in Berlin. The datasets obtained in this study have been uploaded to GEO under accession ID GSE89392.

2.2.1. Acquisition of publicly available datasets for TB multi-cohort analysis

Out of multiple publicly available transcriptomic datasets from blood of TB patients I decided to include 7 studies to create a combined meta-dataset (MDS) and two validation datasets (Table 3).

The presented datasets met the following criteria:

- Contained at least data from untreated TB patients and HCs
- Contained at least 8 samples in each of the groups: TB patients and HC (including LTBI)
- Were performed using platforms which measured at least 18,000 common genes
- Were performed using platforms with annotation available in BiomaRt R package (Durinck et al. 2005, 2009)

Table 3 List of publicly available studies acquired for TB multi-cohort analysis

The accession number refers to the GEO dataset identifier. The study location refers to the cities or countries where the patients were recruited. Number of cases defines the number of TB patients, other disease (OD) patients and healthy (including latently infected) individuals (HC) included in each study.

MDS

Accession number	Citation	Study location	Number of cases
GSE19491	(Berry et al., 2010)	London, SA	54 TB 96 OD 93 HC
GSE47673	(Kaforou et al., 2013)	Malawi, SA	215 TB 194 OD 175 HC
GSE28623	(Maertzdorf, Ota, et al., 2011)	The Gambia	46 TB 62 HC
GSE34608	(Maertzdorf et al., 2012)	Germany	8 TB 18 sarcoidosis 18 HC
GSE42834	(Bloom et al., 2013)	London	35 TB 91 OD 113 HC
GSE39941	(Anderson et al., 2014)	SA, Malawi, Kenya	114 TB 175 OD 57 HC
GSE73408	(Walter et al., 2016)	USA	35 TB 39 pneumonia 35 HC

Validation datasets

Accession number	Citation	Study location	Number of cases
GSE54992	(Cai et al., 2014)	China	9 TB 12 HC
GSE83456	(Blankley, Graham, Turner, et al., 2016)		45 TB 47 EPTB 49 OD 61 HC

The dataset collection selected for the study can be accessed on the website: http://bioinfo.mpiib-berlin.mpg.de/TBprofiles/

2.2.2. Acquisition of publicly available sepsis datasets for the validation of methods

I acquired three publicly available sepsis datasets for the validation of methods used to analyze the TB datasets.

Table 4 List of publicly available studies acquired for sepsis multi-cohort analysis

The accession number refers to the GEO dataset identifier. The study location refers to the cities or countries where the patients were recruited. Number of cases defines the number of sepsis patients, other patients and healthy individuals (HC) included in each study.

Accession number	Citation	Study location	Number of cases
GSE13904	(Wong et al., 2009)	USA	32 sepsis 67 septic Shock 22 SIRS 18HC
GSE9960	(Tang, McLean, Dawes, Huang, & Lin, 2009)	Australia	70 sepsis
GSE28750	(Sutherland et al., 2011)	Australia	27 sepsis 30 post-surgical sepsis 20 HC

2.2.3. Acquisition of GEO datasets for the comparison of mouse and human

This part of the Methods has been adapted from my publication published in September 2017 in Scientific Reports (Domaszewska et al., 2017) and contains fragments related to the experimental procedures (2.2.4, 2.2.5, 2.2.6, 2.2.7,2.2.8) described by my colleagues Lisa Scheuermann and Anca Dorhoi from MPIIB, Department of Immunology, who conducted the experiments. Those listed sub-chapters are identical as in the publication (Domaszewska et al., 2017).

To compare transcriptomic responses against TB in man and mouse I investigated two types of cells: WB cells and macrophages. Blood is a carrier of immune system molecules in the organism and macrophages play a crucial role in the Mtb infection. Therefore, I acquired the following publicly available datasets to compare them with each other as well as with the datasets collected in MPIIB by Lisa Scheuermann and Anca Dorhoi (Department of Immunology) with the help of Karin Hahnke (Department of Immunology) and Hans Mollenkopf (Microarray Core Facility) (Table 5).

Table 5 List of publicly available studies acquired for comparison of human and murine immune response to TB

The column "Dataset name" refers to the names further used in this chapter and in the Chapter 4. The cohort described by Kaforou et al. (2013) contains patients from two geographical locations: SA and Malawi, which have been analyzed separately.

Accession number	Citation	Organism	Sample type	Time points	Number of cases	Dataset name
GSE37250	(Kaforou et al., 2013)	*Homo sapiens*	blood	*N/A*	215 TB 194 OD 175 healthy	SA, Malawi
GSE28623	(Maertzdorf, Ota, et al., 2011)	*Homo sapiens*	blood	*N/A*	46 TB 62 healthy	The Gambia
GSE11199	(Thuong et al., 2008)	*Homo sapiens*	MDM	before infection, 4h p.i.	4 pulmonary TB 4 TB meningitis 4 healthy	GSE11199
GSE47673	(McNab et al., 2011)	*Mus musculus*	BMDM	before infection, 1h p.i., 6h p.i.	4 M.tb. infected 4 uninfected	GSE47673
GSE23508	(Carow et al., 2011)	*Mus musculus*	BMDM	before infection, 24h p.i.	4 M.tb. infected 3 uninfected	GSE23508

2.2.4. Mice and Mtb infection

129S2 (129SvPas) mice were bred and kept under specific pathogen-free (SPF) conditions at the MPIIB in Berlin, Germany. C57BL/6 animals were purchased from Charles Rivers Laboratories. Mice were matched for age and sex and co-housed for at least two weeks under SPF conditions at the MPIIB in Berlin, Germany before start of the experiments. At the time of infection, all mice were 9–12 weeks of age. Aerosol infection with Mtb strain H37Rv and enumeration of bacteria in lung tissue were performed as previously described (Dorhoi et al., 2013). All experiments were approved by the State Office for Health and Social Affairs (Landesamt fuer Gesundheit und Soziales) and conducted in accordance with German Animal Protection Law.

2.2.5. Blood collection and RNA isolation

At indicated time points mice were anesthetized by intraperitoneal injection of 16 mg/kg bodyweight Rompun and 120 mg/kg bodyweight Ketavet in PBS. Blood was drawn from the inferior vena cava of all mice using a 26G needle. 200 µl of blood were directly transferred into 800 µl of TRIzol® (Invitrogen). Total RNA extraction of all blood samples was performed according to the manufacturer's instructions. The RNA yield and A260/280 ratio were measured with a NanoDrop ND 100 spectrometer (NanoDrop Technologies), and RNA integrity was verified using an 2100 Bioanalyzer (Agilent Technologies) with a RNA integrity number (RIN) higher than 7.

2.2.6. *Blood microarrays*

Total RNA of blood samples was labeled with the Low Input Quick Amp Labeling (Agilent Technologies) according to manufacturer's instructions. Quantity and labeling efficiency were verified before hybridization of the samples to SurePrint G3 Mouse GE 8x60K Microarray (Agilent Technologies, Product Number G4852A, Design ID 028005). Scanning of microarrays was performed with 3 μm resolution using a high-resolution laser microarray scanner (Agilent Technologies G2565CA). Quality, reproducibility, and reliability of single microarray data was accessed by the 1-color gene expression QC report from Agilent Technologies.

2.2.7. *Acquisition of THP1 data*

The human monocytic cell line THP-1 (ATCC TIB-202) was maintained in RPMI 1640 (Gibco), supplemented with 10% (v/v) heat-inactivated fetal calf serum (Gibco), 1% (v/v) penicillin–streptomycin (Gibco), 1% (v/v) L-glutamine (Gibco), 1% (v/v) HEPES buffer (Gibco) and 0.05 M 2-mercaptoethanol (Gibco). Cells were differentiated into macrophages by treatment with 50 ng/ml of phorbol 12-myristate 13-acetate (PMA, Calbiochem). Subsequently they were rested for 48 hours and afterwards infected with single-bacterial suspensions of the virulent strain H37Rv, at a multiplicity of infection of 5. At 1, 6 and 24 h following infection, macrophages were lysed with 4M guanidine isothiocyanate solution (Invitrogen), eukaryotic RNA was stabilized in Trizol LS (Invitrogen) and extracted according to vendor's instructions. The RNA yield was detected with a NanoDrop ND 100 spectrometer (NanoDrop Technologies), and RNA integrity was estimated using the 2100 Bioanalyzer (Agilent Technologies).

2.2.8. *Macrophage RNA microarrays*

Total RNA of infected and uninfected THP-1 control cells was labeled with the Quick Amp Labeling (Agilent Technologies) according to manufacturer's instructions. After quality and labeling efficiency control samples were hybridized to 4x44K Whole Human Genome Microarray kits (Agilent Technologies, Product Number G4112F, Design ID 014850). Scanning of microarrays was performed with 5μm resolution using a G2565CA high-resolution laser microarray scanner (Agilent Technologies) using extended dynamic range (XDR). Raw microarray data were extracted with the Agilent FE software V10.5.1.1. and GE1_105_Dec08 protocol using default settings.

2.3. DATA NORMALIZATION

2.3.1. *Data preprocessing*

Data analysis was performed in R version 3.4.3 (2017-11-30), and a script including all analytical steps is available upon request. The datasets have been analyzed with R package *limma* for differential expression analysis (Ritchie et al., 2015). In the studies where raw-datasets were available,

the background signal intensities have been corrected using *normexp* method with offset. The arrays have been quantile normalized. In the mouse-human study, the repeating probes were averaged in the last step of preprocessing which means that the intensities of the probes corresponding to the same gene have been combined. This step has not been made in the multicohort analysis, since the study involved biosignature identification on transcript level with the intention not to miss the transcripts which represent a biological difference, for example in case of alternative splicing. The different approaches result from the assumptions of the two projects: while in the inter-species comparison the conservative approach is preferred due to variability between host genomes, in the multi-cohort analysis I am intending to retain possibly high resolution.

HGNC and ENSEMBL identifiers have been mapped to microarray probe names using biomaRt *mapIds* function (version 2.24.1; Durinck et al., 2005; Durinck, Spellman, Birney, & Huber, 2009).

2.3.2. *Data normalization for multi-cohort analysis*

In the multi-cohort analysis, author-normalization of single studies was used. Each dataset was then divided into training set, containing randomly assigned 80% of the HC (including LTBI), 80% of other disease (OD) and 80% of the TB patients samples (or, in case of the sepsis validation dataset, 80% of the sepsis patient samples), and test set containing the remaining 20% of the samples. MDS was created out of the training sets from each study using only the common genes. Two types of normalization were tested to create the MDS: *ComBat* normalization from Bioconductor *sva* package and standardization based on median and interquartile range (IQR) values, calculated as follows:

$$e'_{i,j} = \frac{e_{i,j} - median\,(e_{.,j})}{IQR_{.,j}}$$

Equation 1

Where:

$e'_{i,j}$ – normalized expression value for gene i,

$e_{i,j}$ – expression measurement of gene i,

$IQR_{.,j}$ – interquartile range for expression measurement of gene i

2.4. DIFFERENTIAL EXPRESSION CALCULATION

Differentially expressed genes between healthy individuals (or uninfected macrophages) and individuals suffering from TB (or Mtb infected macrophages) were identified by creating linear model which included the factors: stimulus type ("TB" and "healthy" or "Mtb infected" and "uninfected") and time point using *lmFit* function from the R package *limma* (Ritchie et al., 2015). The p-values were calculated based on the moderated t-statistic.

In multicohort analysis the differential regulation of genes was calculated after the preprocessing as well as after the normalization described in the chapter 2.3.2. To investigate the influence of normalization on the order of differentially regulated genes GSEA was performed for every dataset before and after the normalization using R package *tmod* (Weiner & Domaszewska, 2016).

2.5. GSEA FOR INDIVIDUAL PATIENTS

To perform GSEA for individual patients, gene expression was expressed as z-score. For every gene, mean expression and standard deviation of gene expression were calculated for healthy individuals from a given cohort. Then, the mean gene expression based on healthy individuals was subtracted from the expression measurement of every individual present in the MDS and the result was divided by standard deviation of gene expression for the healthy patients. The z-score was calculated based on two-sided t-test for the resulting values.

$$e'_{i,j} = \frac{e_{i,j} - mean\,(e_{.,j})}{sd_{.,j}}$$

Equation 2

$$z - score = pnorm(e')$$

Equation 3

Where:

i, j - sample, gene

$e'_{i,j}$ – normalized expression value for gene 'i',

$e_{i,j}$ – measured expression value of gene 'i',

$mean\,(e_{.,j})$– mean measured expression of gene 'i' across all samples,

$sd_{.,j}$ – standard deviation of expression measured for gene 'i' across all samples

GSEA was performed for every individual on the list of genes sorted by increasing z-score using *tmodCERNOtest* function from the R-package *tmod* (Weiner & Domaszewska, 2016) and two sets of gene modules (Chaussabel et al., 2008; Li et al., 2014).

2.6. DEFINITION OF IFN TYPE I AND IFN TYPE II MODULES

The module sets defined by Li et al. (2014) and Chaussabel et al. (2008) are based on the analysis of gene co-expression, in contrast to other frequently used knowledge-based module sets (like GO sets or WikiPathways). It implies that the genes present in the same module might not share the same function or belong to the same signaling pathway. Moreover, a module can consist of a collection of genes the functions of which are only partly known or even completely undefined. For this reason, I decided to implement the knowledge-based distinction into IFN type I and IFN type II induced genes into the modules created by Li et al. and Chaussabel et al. and in each module identified the genes regulated by type I IFN, type II IFN and type I and II IFN signaling pathways together. I referred to Interferome v2.0 database (Rusinova et al., 2012) to specify which of the genes included in the immunological modules are classified as specifically type I IFN response genes, type II IFN response genes or both. The modules DC.M5.12, LI.M158.0 and LI.M158.1 contained mostly genes activated exclusively by type I IFN signaling, and the modules LI.M127, LI.M75, DC.M1.2, DC.M3.4 contained mostly genes activated by both type I and type II IFN signaling pathways. I created three module sets based on the published gene modules created by Li et al. and Chaussabel et al. (Chaussabel et al., 2008; Li et al., 2014) and the classification of genes into IFN type I activated genes, type II IFN activated genes and the genes activated by both type I and type II IFN signaling pathways according to Interferome v2.0 database (Rusinova et al., 2012) (Supplementary Tables 1-4). The created sets consisted of genes which overlapped between originally defined modules and the genes from the MDS classified by Interferome database either as IFN type I inducible genes, IFN type II inducible genes or IFN type I and type II inducible genes. Additional two modules contained (i) all genes classified as IFN type I genes and (ii) all genes classified as IFN type II genes, independent of their original module identity understood as presence in a particular Li et al (2014) or Chaussabel et al. (2008) module set.

2.7. IDENTIFICATION OF IFN+ AND IFN- PATIENTS

GSEA was performed on the list of genes from every individual present in MDS sorted by increasing z-score using the three created module sets. The individuals presenting no significant enrichment in any of the IFN type I modules were classified as "IFN I-". The individuals presenting enrichment in at least one IFN type I module were classified as "IFN I+". Similarly, the "IFN II$^-$" and "IFN I and II-" individuals presented no enrichment in the IFN type II or IFN type I and II module set respectively, and those presenting enrichment were classified as "IFN II+" or "IFN I and II+". Ultimately, the overlap between "IFN I+", "IFN II+" and "IFN I and II+" patients was analyzed and compared to classification to "IFN+" and "IFN-" patients based on the original sets of modules published by Li et al. (2014) and Chaussabel et al. (2008).

2.8. LOGISTIC REGRESSION

Each sample in MDS was characterized with the following criteria: study, ethnicity, residence, nationality, TB status, HIV status, other detected diseases and IFN status. Two logistic regression models were fit using *glm* function from R package *stats* (R Core Team, 2018), with IFN status as dependent variable and TB status, study, ethnicity, residence, HIV and OD as predictor variables using whole MDS (i), and study, ethnicity, residence, HIV and OD as predictor variables using only the subset of data characterized by active TB (ii).

2.9. IDENTIFICATION OF CONCORDANT AND DISCORDANT GENES BETWEEN IFN+ AND IFN- TB PATIENTS

I calculated differential expression of genes between TB IFN+ patients and healthy and TB IFN- patients and healthy. Subsequently, I used R package *disco* to calculate disco.score expressing the concordance and discordance of gene regulation to identify the concordantly and discordantly regulated genes between IFN+ and IFN- patients.

2.10. CYTOKINE CONCENTRATIONS IN BLOOD OF IFN I$^+$ AND IFN I$^-$ INDIVIDUALS

Mtb infection is one of the many known stimulations of IFN response. To validate the approach of dividing people into IFN- and IFN+ individuals I searched for other stimuli inducing IFN response and datasets containing both transcriptomic studies of the WB of the patients with such stimulations as well as measurements of the actual levels of the IFN-inducible cytokines in blood of those patients. Other factors inducing IFN response include for example influenza virus infection. The vaccines against influenza cause strong IFN response in humans (Athale et al., 2017; Banzhoff et al.,

2008). Datasets collected within BioVacSafe project kindly shared with me by Jeroen Maertzdorf and January Weiner were derived from blood of 114 healthy volunteers who underwent FLUAD™ vaccination. I used the datasets to compare GSEA-based IFN I-status definition with the blood absolute concentrations of cytokines CXCL10 and CCL2, which are inducible by IFN type I signaling. Blood samples were collected before the vaccination on the vaccination day and at day 1 after the vaccination. The samples were used to perform microarrays and cytokine measurements. The dataset is available upon request.

GSEA was calculated using the IFN type I module set and each sample was assigned IFN I+ or IFN I- status as described before. The fold change between absolute blood concentrations of cytokines CXCL10 and CCL2 in day 1 post vaccination and before the vaccination were calculated for the IFN I+ and IFN I- samples.

2.11. CORRELATION BETWEEN IFN STATUS AND DISEASE SEVERITY

The dataset GSE19491 (Berry et al., 2010) was used to compare the IFN status with the disease severity assessed on the basis of lung X-Ray studies of TB patients and HCs. The IFN type I status was assessed using microarray results and GSEA as described above for all the participants of the study (61 TB patients, 105 HCs including 69 LTBI, and 274 OD patients). 72 individuals from the study cohort underwent the lung X-Ray investigation and were diagnosed as "healthy" (n = 34), "minimal disease" (n = 14), "moderate disease" (n = 13), or "advanced disease" (n = 11) by doctors blinded to the microarray results and the clinical diagnosis of the patients. The X-Ray based diagnosis was compared with the IFN I status calculated using GSEA.

2.12. MACHINE LEARNING METHODS

2.12.1. Unsupervised Machine Learning - PCA

PCA was performed on the MDS as well as on the subset of MDS containing only samples from active TB patients using R-packages *stats*, *pca3d* and *tmod* (R Core Team, 2018; Weiner, 2017; Weiner & Domaszewska, 2016). The fraction of variance explained by each factorial predictor (including TB status, IFN status, study, ethnicity, residence, HIV, OD, microarray platform) was calculated for each principal component using *prcomp* function from the *stats* package which performs singular value decomposition of the centered and scaled data matrix (R Core Team, 2018). Among the first 11 PCs I searched for two PCs explaining the highest proportion of variance for each factor detected as significant by logistic regression (described in the chapter 2.8) using 100-fold randomization. The two PCs with the highest fit (r^2 – coefficient of determination) were chosen. The randomization as well

as data distribution along the two chosen PCs was illustrated. GSEA for PCs was calculated on the genes ranked by their weights in a given PC and visualized using R-package *tmod*.

2.12.2. Supervised Machine Learning - Random Forest models

Random Forest models with 10-fold cross validation

RF models were created to classify patients of TB with or without IFN response and (i) healthy, (ii) OD, (iii) non-TB (containing both healthy and OD) patients as shown in Table 8 in the Chapter 3. The appropriate subsets of the training MDS containing TB patients with or without IFN response and (i) HCs, (ii) OD patients, (iii) non-TB patients were selected and split into 10 folds. Class balancing was used to retain the proportion of one case to three control individuals. The RF models were repetitively trained on the data from 9 folds and tested on the remaining fold. Performance of the models was evaluated by creating receiver-operator characteristic (ROC) curves.

Additional six RF models were created in analogical way with exclusion of genes involved in IFN I signaling to investigate their influence on the RF results (Table 8).

The RF models were created using R package *randomForest* (Kuhn, 2008; Liaw & Wiener, 2002) and cross-validation was performed using R package *caret* (Kuhn, 2008). The ROC plots were created using R package *pROC* (Robin et al., 2011).

Determination of the signature size

To determine the biosignatures of IFN- and IFN+ TB patients I used the models 5 and 6 described in the Table 8 in the Chapter 3 since those models identified TB patients among all other patient groups – HC, LTBI and OD. To determine the minimal number of transcripts required to discriminate TB from other patient groups I sorted the transcripts in both models by decreasing variable importance. I defined the TB IFN+ and TB IFN- signatures consisting of top (i) 5, (ii) 7, (iii) 10, (iv) 20, (v) 50 or (vi) 200 ranking transcripts, and created new models trained only with use of the selected transcripts. I then tested the new models using 10-fold cross validation within the training MDS and evaluated the performance of the models using ROC plot. The size of biosignature giving AUC higher than 0.8 with possible small number of included transcripts was chosen as the optimal biosignature size.

Determination of the TB IFN+ and TB IFN- biosignatures

For the detection of TB IFN+ and TB IFN- biosignatures two new RF models with class balancing retaining the proportion of one TB to three non-TB cases were trained using the subsets of the complete training MDS subsets containing (i) all TB IFN+ and non-TB (Biosignature Model 1), (ii) all TB IFN- and non-TB (Biosignature Model 2). A biosignature of 20 top ranking transcripts from the Biosignature Model 1 sorted by variable importance was defined as IFN+ biosignature. A biosignature

of 50 top ranking transcripts from the Biosignature Model 2 sorted by variable importance was defined as IFN- biosignature.

Testing of the TB IFN+ and TB IFN- biosignatures

The obtained TB IFN+ and TB IFN- biosignatures were tested on the test MDS and their performance has been evaluated using ROC.

Validation of the TB IFN+ and TB IFN- biosignatures

The obtained TB IFN+ and TB IFN- biosignatures were tested on the external dataset from Cai et al. and their performance has been evaluated using ROC.

2.13. VALIDATION OF THE SIGNATURE FINDING PIPELINE ON SEPSIS META-DATASET

Sepsis training and test MDS were created following the steps described in the chapters 2.2 and 2.3. GSEA was calculated for individual patients in the sepsis training and test MDS and the IFN status was assigned to the individual donors as described in the chapters 2.5 and 2.7. For the detection of sepsis IFN+ and sepsis IFN- biosignatures two RF models with class balancing retaining the proportion of one sepsis to three HC cases were trained using the subsets of the complete training sepsis MDS containing (i) sepsis IFN+ and healthy (Sepsis Model 1), (ii) sepsis IFN- and healthy (Sepsis Model 2). Biosignatures of 10 top ranking transcripts from the Sepsis Model 1 and Sepsis Model 2 sorted by variable importance were defined as IFN+ sepsis biosignature and IFN- sepsis biosignature, correspondingly. The obtained sepsis IFN+ and sepsis IFN- biosignatures were tested on the sepsis test MDS and their performance has been evaluated using ROC. The transcripts present in the biosignatures were compared with previously published sepsis biosignatures.

2.14. CORRELATION MATRIX

Gene sets like the above described BTMs capture the relationship between transcripts measured on the microarrays. When annotated, they suggest biological interpretation of activation programs launched by cells in response to stimulation. From the systemic point of view, the co-expression modules are components of a meta-network which reveals a higher-level organization of the transcriptome (Langfelder & Horvath, 2008). In other words, understanding transcript-organization in modules is a preliminary step to understand the network of module interactions. Exploring this network can be executed thanks to finding module representatives. In case of modules containing groups of genes with correlated expression such a representative is the eigengene, which summarizes the module expression profile (Langfelder & Horvath, 2008). Such defined representatives or eigengenes of BTMs can be correlated creating a correlation matrix informative of relationships between the different

modules (Langfelder & Horvath, 2008). For example, as further presented in this thesis, we can learn that upregulation of gene expression in a particular module can be correlated with downregulation of the expression of genes present in another module.

The framework of correlation network analysis applied in this study consists of the following steps:

- Identification of the most significantly enriched gene modules among TB patients in comparison with healthy
- Calculation of eigengene of each module
- Calculation of correlation matrix between the eigengenes
- Visualization of the correlation matrix using clustering according to the correlation level.

The gene modules presenting enrichment in GSEA in (i) at least one TB patient with p-value lower than 10^{-11} and AUC of at least 0.85 (in case of the human MDS), or (ii) at least one Mtb infected macaque with p-value lower than 10^{-3} and AUC of at least 0.7 (in case of the macaque dataset) were selected for the correlation calculation. The eigengenes of every module as well as the Pearson correlation between the expressions of genes in every module were calculated using the function *cormods* from R package *tmod* (Weiner & Domaszewska, 2016) including only the genes which expression was measured across all samples. Hierarchical clustering was employed to cluster the enriched modules according to eigengene correlation using the functions *dist* and *hclust* from R package *stats* (R Core Team, 2018). The function *dist* calculates distances between vectors which are rows of expression matrix and the function *hclust* performs hierarchical clustering on the calculated distances by initially assigning objects to their own clusters and then iteratively joining the most similar clusters with Ward's minimum variance method (Ward, 1963).

The correlation of gene expression in the clustered modules was visualized using *ggplot2* R package (Wickham, 2009).

2.15. DISEASE RISK SCORE APPLICATION

I used the method described by Kaforou et al. (2013) for converting complex multiple transcript RNA signatures to obtain disease risk score (DRS) which is a single value score for every individual. I calculated the DRS on the basis of TB signatures (Kaforou et al., 2013) for three setups: (i) TB patients and healthy, on the basis of 27-gene TB vs healthy signature, (ii) TB patients and OD patients, on the basis of 44- gene TB vs OD signature, and (iii) TB patients vs non-TB, on the basis of 53-gene TB vs non-TB signature. Depending on the setup, the classification of TB/healthy/OD/non-TB status on the basis of DRS calculation was assigned to every individual and compared with the IFN status.

2.16. INFLUENCE OF TIME POST INFECTION ON INTERFERON STATUS

To investigate if the IFN status in individuals with active TB is the result of time p.i., a longitudinal dataset generated to assess changes in WB gene expression after Mtb infection in Cynomolgus macaques was acquired from GEO database (GSE84152; Gideon et al., 2016). The dataset contained microarray results collected from 38 macaques at two time points before Mtb infection and at days 3, 7, 10, 20, 30, 42, 56, 90, 120, 150, 180 p.i.. The samples were normalized and z-score was calculated using the aforementioned method (chapter 2.3, 2.5). GSEA was performed on samples from individual macaques. The samples were assigned IFN I+/IFN I- status which was compared with their binary clinical diagnosis and lung inflammation.

2.17. ORTHOLOGS ASSIGNMENT BETWEEN HUMAN AND MURINE DATASETS

A table of expected comparisons between murine and human samples was created (Table 6). Orthologous genes were assigned to each other between corresponding human and mouse datasets used in each comparison. Probe names specific to the microarray used were assigned an ENSEMBL identifier with use of "mapIds" function from biomaRt package (version 2.24.1; Durinck et al., 2005, 2009). Then, orthologous human and mouse genes were identified with biomaRt *getLDS* function based on homology mapping between different species interlinked in Ensembl database (with attributes and filters defined as "ensembl_gene_id"). Only the putative orthologs with a 1:1 mapping (no potential in-paralogs) were included in the further analysis.

Table 6 List of the comparisons performed on the human and murine datasets

"Comparison ID" is further used in the Chapter 4. "Human dataset" column refers to (i) the cohort origin as previously presented in the Table 5 (The Gambia, SA, Malawi) in case of publicly available human WB datasets, (ii) datasets from THP1 cells acquired in the MPIIB (THP1) and (iii) datasets ID from GEO in case of the publicly available human macrophage datasets. The column "Murine dataset" refers to (i) datasets from murine WB acquired at MPIIB ("C57BL/6" and "129S2"), (ii) datasets ID from GEO in case of the publicly available murine macrophage datasets. *N/A* – not applicable.

Comparison ID	Human dataset name	Human tissue	Human time point	Murine dataset name	Murine tissue	Murine time point
1	The Gambia	blood	*N/A*	C57BL/6	blood	Day 1
2	The Gambia	blood	*N/A*	129S2	blood	Day 1
3	Malawi	blood	*N/A*	C57BL/6	blood	Day 1
4	Malawi	blood	*N/A*	129S2	blood	Day 1
5	SA	blood	*N/A*	C57BL/6	blood	Day 1
6	SA	blood	*N/A*	129S2	blood	Day 1

7	The Gambia	blood	*N/A*	C57BL/6	blood	Day 7
8	The Gambia	blood	*N/A*	129S2	blood	Day 7
9	Malawi	blood	*N/A*	C57BL/6	blood	Day 7
10	Malawi	blood	*N/A*	129S2	blood	Day 7
11	SA	blood	*N/A*	C57BL/6	blood	Day 7
12	SA	blood	*N/A*	129S2	blood	Day 7
13	The Gambia	blood	*N/A*	C57BL/6	blood	Day 14
14	The Gambia	blood	*N/A*	129S2	blood	Day 14
15	Malawi	blood	*N/A*	C57BL/6	blood	Day 14
16	Malawi	blood	*N/A*	129S2	blood	Day 14
17	SA	blood	*N/A*	C57BL/6	blood	Day 14
18	SA	blood	*N/A*	129S2	blood	Day 14
19	The Gambia	blood	*N/A*	C57BL/6	blood	Day 21
20	The Gambia	blood	*N/A*	129S2	blood	Day 21
21	Malawi	blood	*N/A*	C57BL/6	blood	Day 21
22	Malawi	blood	*N/A*	129S2	blood	Day 21
23	SA	blood	*N/A*	C57BL/6	blood	Day 21
24	SA	blood	*N/A*	129S2	blood	Day 21
25	THP1	THP1	24h	GSE23508	BMDM	24h
26	THP1	THP1	6h	GSE47673	BMDM	6h
27	GSE11199	MDM	4h	GSE23508	BMDM	24h
28	GSE11199	MDM	4h	GSE47673	BMDM	6h
29	GSE11199	MDM	4h	GSE23508	BMDM	24h
30	GSE11199	MDM	4h	GSE47673	BMDM	6h
31	GSE11199	MDM	4h	GSE23508	BMDM	24h
32	GSE11199	MDM	4h	GSE47673	BMDM	6h
33	THP1	THP1	1h	GSE47673	BMDM	1h
34	GSE11199	MDM	4h	GSE47673	BMDM	1h
35	GSE11199	MDM	4h	GSE47673	BMDM	1h
36	GSE11199	MDM	4h	GSE47673	BMDM	1h

2.18. DISCO.SCORE CALCULATION AND GENE SET ENRICHMENT ANALYSIS

I have created an R-package *disco* for identification and visualization of concordant and discordant gene modules. The disco package is available on CRAN (http://cran.r-project.org/web/packages/disco/). Disco score was calculated for each pair of orthologous genes using *discoScore* function. Concordantly and discordantly regulated gene sets were identified by performing GSEA with R-package *tmod* (version 0.27; Weiner & Domaszewska, 2016) using CERNO statistical test, which is a variant of Fisher's method adapted to GSEA (Yamaguchi et al., 2008) on the list of genes sorted by the decreasing or increasing disco.score, respectively. Disco.score for particular genes has been visualized with the color gradient on the plots presenting $\log_2$ of fold change ($\log_2$FC) of gene expression in stimulated vs non-stimulated organisms.

The general formula for disco score applicable to n datasets is defined by the equation 4:

$$disco.score = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} log_2 FC_i \cdot log_2 FC_j \cdot (log_{10} P_i + log_{10} P_j)$$

Equation 4

where:

n – number of datasets analyzed

FC_i - fold change for gene in the dataset i, as calculated in differential expression analysis

FC_j - fold change for gene in the dataset j, as calculated in differential expression analysis

P_i - p-value for human gene in the dataset i, as calculated in differential expression analysis

P_j - p-value for murine gene in the dataset j, as calculated in differential expression analysis

In this thesis I use the formula for disco score applicable to two heterologous datasets, e.g. one from human patients and one from mice, defined by the equation 5:

$$disco.score = log_2 FC_{Hs} \cdot log_2 FC_{Mm} \cdot |log_{10} P_{Hs} + log_{10} P_{Mm}|$$

Equation 5

where:

FC_{Hs} - fold change for gene in the human dataset, as calculated in differential expression analysis

FC_{Mm} - fold change for gene in the murine dataset, as calculated in differential expression analysis

P_{Hs} - p-value for human gene in the human dataset, as calculated in differential expression analysis

P_{Mm} - p-value for murine gene in the murine dataset, as calculated in differential expression analysis

2.19. VALIDATION OF DISCO.SCORE WITH SIMULATED MODULES

I have used the human dataset from The Gambia and mouse dataset 21 days p.i. (129S2 mice) and a simulated set of modules to test the performance of disco.score algorithm in retrieving concordantly and discordantly regulated modules. I used the existing murine and human datasets, but I have simulated the assignment of genes to gene sets, thus defining a priori which gene sets contain concordant genes, which gene sets contain discordant genes, and which are negative controls. I then tested whether the disco.score algorithm is able to detect these a priori defined gene sets.

I have simulated gene sets containing 10, 20, 30, 40 or 50 genes, out of which 10%, 20% or 30% were either concordantly regulated or discordantly regulated. In addition, I have generated modules consisting of 10 to 50 genes, which contained equal number of either concordantly or discordantly regulated genes. Each parameter combination (number of genes, number of regulated genes, type of regulation: concordant, discordant or equal number) has been replicated 100 times; an equal number of 100 replicates of a suitable negative control modules was then added to the superset. For concordant modules, the control modules contained only non-concordant genes (including discordant genes, non-regulated genes or genes with significant differences only in one organism); for discordant, only non-discordant genes; for equal number, only genes that were neither concordant nor discordant.

Next, with each set of 200 modules (out of which 100 were concordant or discordant and 100 were negative controls) I performed CERNO test on the list of genes sorted by disco.score and identified the concordant and discordant modules. Then, I sorted the detected modules according to the p-values for enrichment and calculated area under curve (AUC) that corresponds to how accurately the algorithm detected the concordant or discordant modules.

2.20. POSITIVE CONTROLS

I used the dataset from Maertzdorf et al., 2011 containing WB expression profiles from patients suffering from TB, from sarcoidosis and HCs as positive controls for disco.score. The two diseases give expression profiles indistinguishable from each other when compared to HCs. I calculated differential expression between the 18 sarcoidosis patients and 18 HCs and between the 8 TB patients and 18 HCs. I matched the genes between both groups and calculated disco.score for each pair of corresponding human genes. I then sorted the list of differentially expressed genes by decreasing disco.score and performed GSEA to distinguish concordant gene modules between the two groups. Then I sorted the gene list according to increasing value of the disco.score and performed GSEA to distinguish discordant gene modules.

3. CHAPTER 3: EXPLORATION OF INDIVIDUAL VARIABILITY IN HOST RESPONSE TO TUBERCULOSIS

In this section I describe the outcome of the multi-cohort analysis of gene expression regulation in individual TB patients. I show how the conducted analysis brought the focus to individual variability in IFN response to TB among patients.

The analysis of a large meta-dataset composed of 7 publicly available datasets is presented starting from a demonstration of how the choice of normalization method influences outcome of multi-cohort study. Further, I guide the reader through a thorough analysis of the non-TB related factors which, if not accounted for, can affect the results of the analysis and occlude meaningful conclusions. Comparison of the TB patients and healthy individuals on the level of an individual indicates that there are patients who do not present typical for TB regulation in IFN signaling genes. I investigated how does the gene expression of those individuals correspond with the scale of the disease and prove the phenomenon of individual variability in the scale of presented IFN response using datasets where healthy individuals' immune response was triggered by vaccination. The findings in human cohorts are compared to controlled animal TB studies. Finally, I present different biosignatures that characterize the subgroups of patients with strongly pronounced in contrast to insignificant IFN type I response and present the identified patterns of response to TB.

3.1. ABSTRACT

In the last years significant progress has been made in the understanding of immunity against TB. To a large extent it is owed to the multiple published transcriptomic studies of TB patients across different geographical locations. Apart from the single-cohort studies, multi-cohort approaches integrated the publicly available datasets. Both the single and the multi-cohort analyses investigated the trends presented by gene expression regulation in TB patients *vs* non-TB and indicated gene signatures of TB. Various mechanisms dominating the immune response to TB including IFN and complement system signaling have previously been proposed.

Here, I inspect the transcriptomic response to Mtb infection in individual TB patients WB samples collected from the published studies. To investigate individual variability I suggest a method of integration of the author-normalized data and transformation of the measured gene expression levels into z-score. I create a meta-dataset consisting of seven previously published transcriptomic studies from various geographical locations. This is followed by GSEA for individual patients by using two sets of previously published immunological gene modules and novel sets of modules related specifically to type I or type II IFN signaling. The obtained enrichment profiles revealed different patterns of immune response regulation in TB patients, in particular in the IFN responses of the individuals from every cohort. I further investigate the variability in the enrichment of IFN modules and assign the patients into one of the two groups: presenting or lacking the enrichment in IFN-related modules (IFN+ or IFN-).

The variability between those two groups could not be fully explained by any of the factors characterizing the patients which I show using logistic regression models and unsupervised analysis. The division was reflected by higher levels of IFN-inducible cytokines in the WB of the IFN+ when compared to IFN- individuals and corresponded with the severity of the lung pathology of TB patients. ML models for classification of the TB patients among non-TB created based on IFN+ or IFN- TB patients profiles differed in size, composition and their performance in detecting TB patients in the test and validation sets. While the IFN+ TB biosignature presented very unstable behavior detecting TB IFN- individuals, the TB IFN- signature was characterized by stable performance and similar overall sensitivity and specificity of detecting TB IFN+ as well as TB IFN- among non-TB patients.

I suggest that the response to TB is highly variable and dependent on the host and present six gene expression patterns characterizing different subgroups of the patients rather than, as suggested before, a universal transcriptomic response pattern and gene signature of all TB patients.

3.2. DATA ACQUISITION

Seven datasets were analyzed in this study. To create a meta-dataset (MDS) representing possibly variable cohorts, studies conducted in 7 different geographical locations: UK, Germany, SA, The Gambia, Malawi, USA and Kenya were included. In the first step, participants undergoing longitudinal studies and participants which were undergoing anti-TB therapy were excluded since I did not intend to investigate treatment progress, which resulted in the total count of 1959 individuals included in the analysis, with 570 active TB patients, 827 OD patients and 562 healthy donors. 80% of the donors from each of the disease groups: TB, healthy (including uninfected as well as LTBI and HIV+ donors) and OD were randomly selected and collected as training MDS on which all the further analysis steps except for testing the RF models were performed. The other 20% of the donors were assigned to the test MDS. I created a table of metadata describing study participants, which included donor identifier as assigned on GEO, study affiliation, ethnicity, residence, TB status, HIV status and other known diseases according to the descriptions published on GEO database and in the corresponding publications to be able to include those factors in the further analysis (Table 7).

Table 7 Example fragment of the created meta-data table

The full meta-data table is available on the website: http://bioinfo.mpiib-berlin.mpg.de/TBprofiles/. The columns present the donor ID from the GEO database, reference to the original publication, ethnicity of the patients, country of residence, TB status, LTB infection, HIV infection and presence of other disease (OD).

donor	study	ethnicity	residence	TB	LTBI	HIV	OD
GSM851868	(Maertzdorf et al., 2012)	European	Germany	yes	no	no	no
GSM709468	(Maertzdorf, Ota, et al., 2011)	African	The Gambia	no	yes	no	no
GSM914451	(Kaforou et al., 2013)	African	Malawi	yes	no	yes	no
GSM1050945	(Bloom et al., 2013)	Asian	Europe	no	no	no	pneumonia

As a result, the training and test MDS were created which consisted of the samples from donors from 7 cohorts, including (i) 452 TB donors, 665 OD donors and 457 healthy donors in the training MDS and (ii) 110 TB donors, 162 OD donors and 113 healthy donors in the test MDS, which was at that moment left out of the analysis.

3.3. DATA NORMALIZATION

Data normalization has been performed separately for the training and test MDS. For each study in the MDS I calculated differential gene expression between TB patients and HCs. Since technical variation and batch effects play an important role in microarray-based transcriptome studies I tested two normalization procedures which should minimize their effects. The first, ComBat, is

implemented in the Bioconductor package *sva* and combines mean and variance adjustment for each batch and every gene separately with empirical Bayes normalization (Leek et al., 2018). In the second approach I made use of the fact that the median and interquartile range (IQR) of the expression of every gene are known and standardized their values between different studies according to Equation 1 in chapter 2.3.2.

The data distribution after normalization with ComBat presented less variance than the data distribution after the median and IQR-based type of normalization (Figure 5).

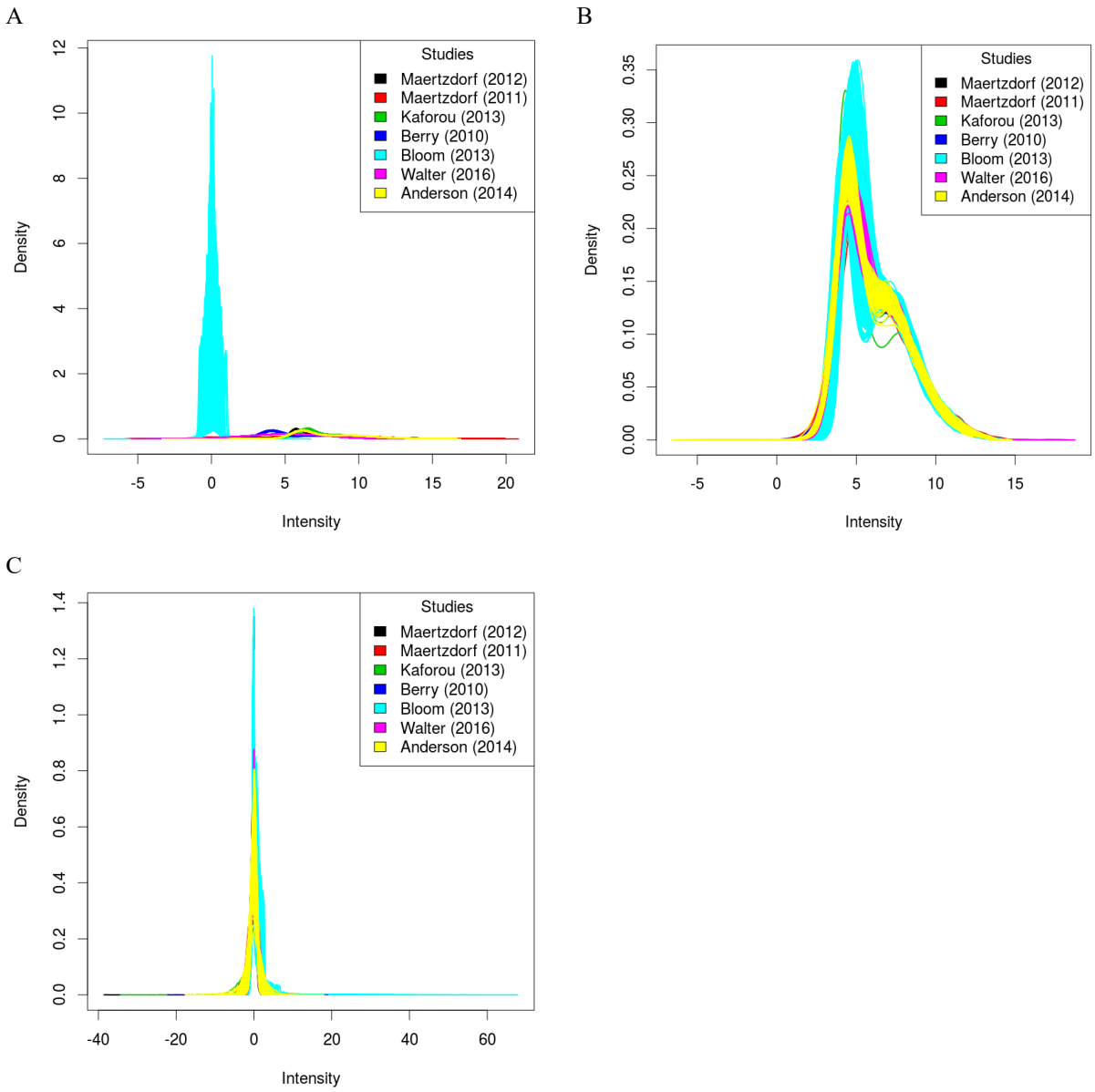


Figure 5 Distribution of data in MDS before and after the tested normalizations

(A) Datasets before normalization, (B) after ComBat normalization, (C) after normalization with median and IQR. The datasets are colored by study. The best normalization of data distribution is achieved with ComBat data normalization procedure.

Subsequently I applied GSEA to the lists of genes sorted by increasing p-values calculated for differential regulation in every study before normalization and after both normalization methods (Figure 6). The resulting enrichments and density plots for all the datasets before and after normalization indicated that although the second method results in more variance in data distribution, it preserves the order of the genes when sorted by differential regulation, while the ComBat normalization significantly changed the gene order and reduced the enrichment. Therefore, this normalization reduced the relevant biological information contained in the separate studies compromising it for the sake of better variance reduction.

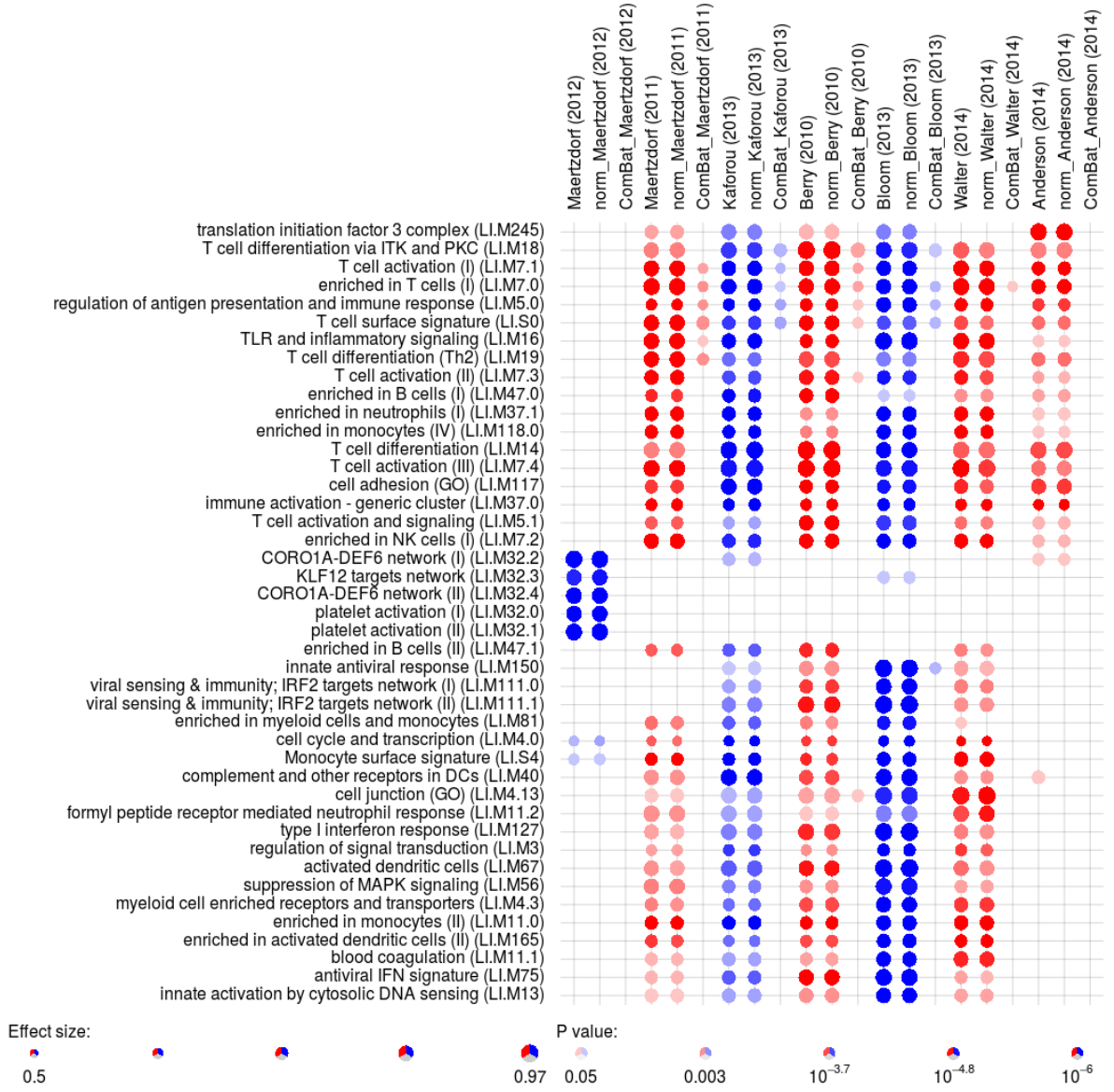


Figure 6 GSEA performed on all the studies before and after the two tested normalization methods

The columns are named after the study author. The column names of the columns presenting enrichment after the median and IQR normalization start with "norm_" and the column names of the columns presenting enrichment after Combat normaliyation start with "ComBat_"). The dots present enrichment in the modules described in the row names. The dot size is proportional to effect size and the intensity of the color is proportional to decrease in p-value of the enrichment. The red and blue colors are used only to facilitate distinguishing subsequent studies.

Since the median and IQR- based standardization procedure did not influence the gene order and at the same time standardized the data, I used these calculated standardized expression values in further analysis. For each standardized expression value I calculated the z-score according to Equation 2 and Equation 3 (chapter 2.3.2).

3.4. GENE SET ENRICHMENT ANALYSIS

To identify mechanisms activated in TB patients in response to the disease I performed GSEA for individual patients on the lists of genes sorted by increasing z-score (Figure 7).

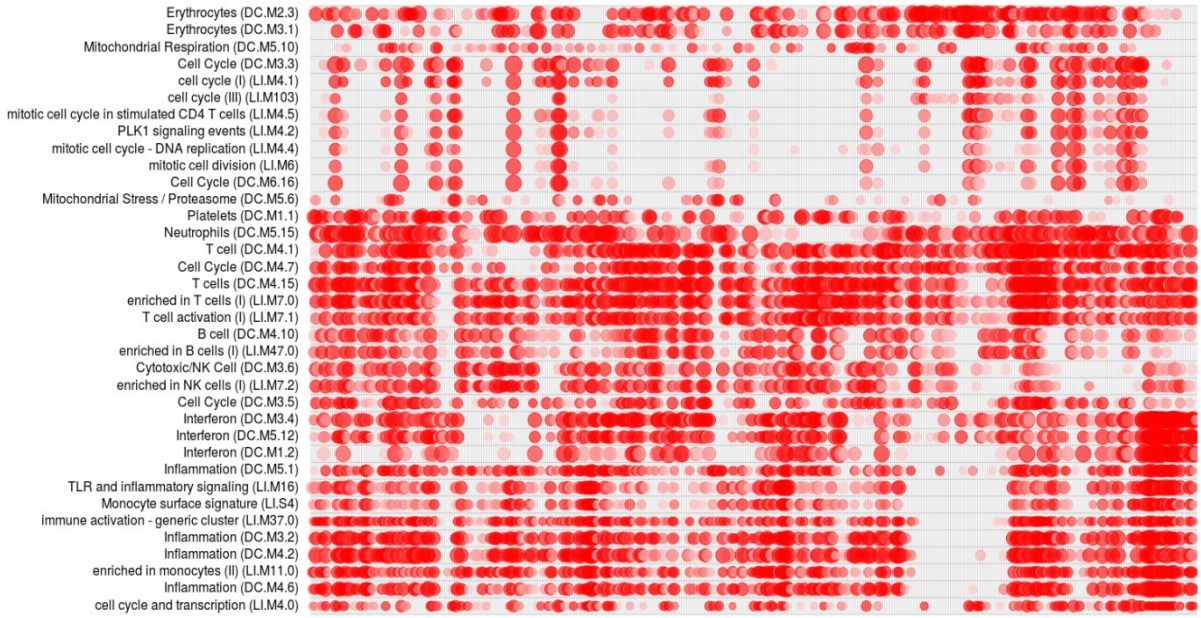


Figure 7 GSEA results for individual patients with TB present in MDS

The single patient profiles are represented in columns. The red dots present enrichment in the modules described in the row names. The dot size is proportional to effect size and the intensity of the color is proportional to decrease in p-value of the enrichment.

GSEA for individual patients presented marked differences in immune responses between TB patients. Strikingly, IFN related modules as well as many other modules considered characteristic in TB patients were not uniformly enriched among the individuals (as shown in the sample of MDS presented in Figure 8). In every cohort, individuals without significant enrichment in IFN modules were present. There was a visible variability in immune responses to TB among individuals from the training MDS. I have shown that various patients present activation of different elements of immune system in response to TB on the gene expression level.

Figure 8 GSEA results for selected TB patients from every cohort

Random groups of patients have been selected from every study cohort to show variability between patients within the cohorts. The single patient profiles are represented in columns. Column names correspond to the GEO ID of the study as described in the Table 3. The dots represent enrichment in the modules described in the row names. The dot size is proportional to effect size and the intensity of the color is proportional to decrease in p-value of the enrichment.

3.5. DEFINITION OF TYPE I AND TYPE II INTERFERON MODULES

For the primary enrichment testing I used the gene modules published by Li et al. (2014) and Chaussabel et al. (2008) among which altogether 7 modules were related to IFN signaling. However, a closer inspection of the genes in each of the 7 modules revealed that two of them contain mostly genes related to IFN-α signaling and are not enriched among TB patients, whereas all the remaining 5 modules presenting significant enrichment contain a mixture of IFN type I and type II signaling genes (Supplementary Table 1). As described in the introduction, chapter 1.3.3, the activation of type I and type II IFN signaling pathways results in dramatically different effects for TB patients, namely increased pathology in case of IFN type I signaling and control of infection in case of IFN type II signaling, therefore distinguishing these two types of response is crucial.

To amend this I referred to the Interferome v2.0 database (Rusinova et al., 2012) which contains vast collection of manually curated publicly available microarray datasets examining IFN type I, II and III responses to various stimulations. I used the Interferome v2.0 database to select IFN-related genes from all the genes present in MDS and classified them as inducible by IFN type I exclusively, IFN type II exclusively or by both types of IFN. Hence, I created three novel sets of modules:

(i) related to IFN type I signaling, consisting of subsets of the original Li et al. (2014) and Chaussabel et al. (2008) modules which overlapped with the collection of IFN type I genes identified by Interferome v2.0 database and one module gathering all genes identified by the Interferome database as IFN I inducible (Supplementary Tables 1 and 2),

(ii) related to IFN type II signaling, consisting of subsets of the original Li et al. (2014) and Chaussabel et al. (2008) modules which overlapped with the collection of IFN type II genes identified by Interferome v2.0 database and one module gathering all genes identified by the database as IFN II inducible (Supplementary Tables 1 and 3),

(iii) related to both type I and type II IFN signaling pathways, consisting of subsets of the original Li et al. (2014) and Chaussabel et al. (2008) modules which overlapped with the collection of IFN type I and II genes identified by Interferome v2.0 database and one module gathering all genes identified by the database as both IFN I and IFN II inducible (Supplementary Tables 1 and 4).

3.6. IDENTIFICATION OF IFN+ AND IFN- PATIENTS

To find individuals with strong IFN response to TB and examine, what types of IFN response are dominating in the patients, I subsequently performed enrichment testing using the three created module sets: IFN I module set, IFN II module set and IFN I and II module set. The donors presenting significant enrichment in the IFN I module set were classified as IFN I+, the donors presenting enrichment in the IFN II module set as IFN II+ and the donors presenting enrichment in the IFN type I and II module set as IFN I and II+. Out of 457 TB patients present in the MDS, 320 presented with enrichment in at least one of the modules from IFN type I module set, 293 in IFN type II and 314 in IFN type I and II module set. Altogether 234 TB patients presented enrichment in all three IFN module sets and 380 TB patients in any of them, indicating high degree of overlap between the patients which present with IFN type I and IFN type II responses (Figure 9). Therefore, I identified subgroups of TB patients which presented IFN response and showed that IFN type I and type II responses are coexisting in the majority of TB patients.

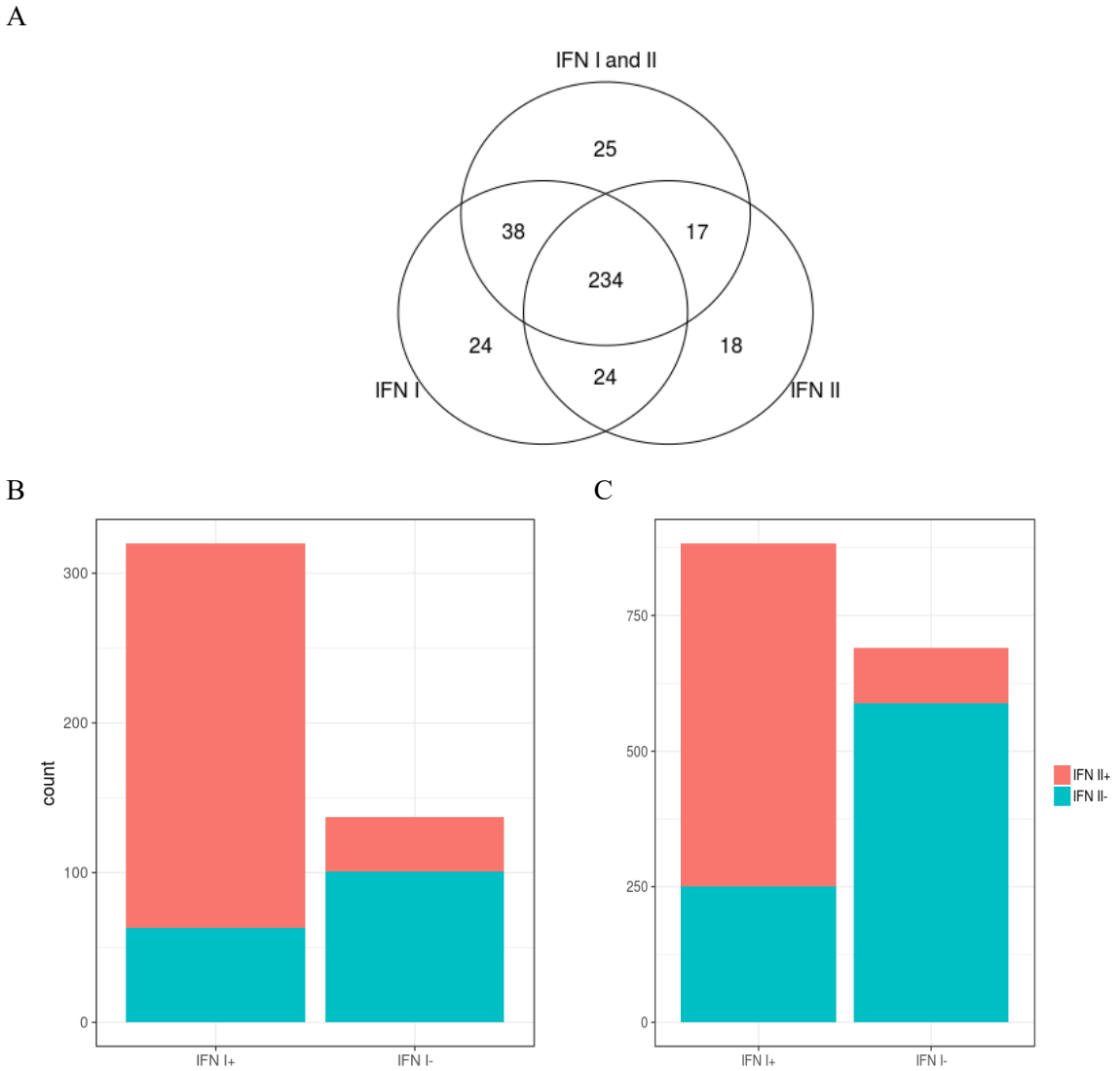


Figure 9 Individuals presenting enrichment in IFN I and IFN II modules

(A) Numbers of patients presenting enrichment in general IFN modules defined by Li et al. (2014) and Chaussabel et al. (2008) (set "IFN") and in IFN type I set ("IFN I"), IFN type II set ("IFN II") and set with genes common for IFN type I and type II signaling ("IFN I and II") defined in the section (3.5). (B) Numbers of individuals with IFN type II enrichment among individuals with and without IFN type I enrichment. (C) Numbers of TB patients with IFN type II enrichment among individuals with and without IFN type I enrichment.

3.7. LOGISTIC REGRESSION AND PRINCIPAL COMPONENT ANALYSIS

To examine the possibility, that IFN status can be influenced by one of the factors describing the included studies or characteristics of the patients I investigated the influence on the presence of IFN I response of the following factors: study, used microarray platform, ethnicity, residence, other diseases, HIV and TB status using ANOVA and logistic regression (GLM). Active TB, HIV and *Streptococcus sp.* and *Staphylococcus sp.* coinfection turned out to be significant predictors of IFN I+ status (Figure 10). The IFN+ status of around 50% of patients without TB results from the fact that the "no TB" group also included patients with OD and HIV coinfection.

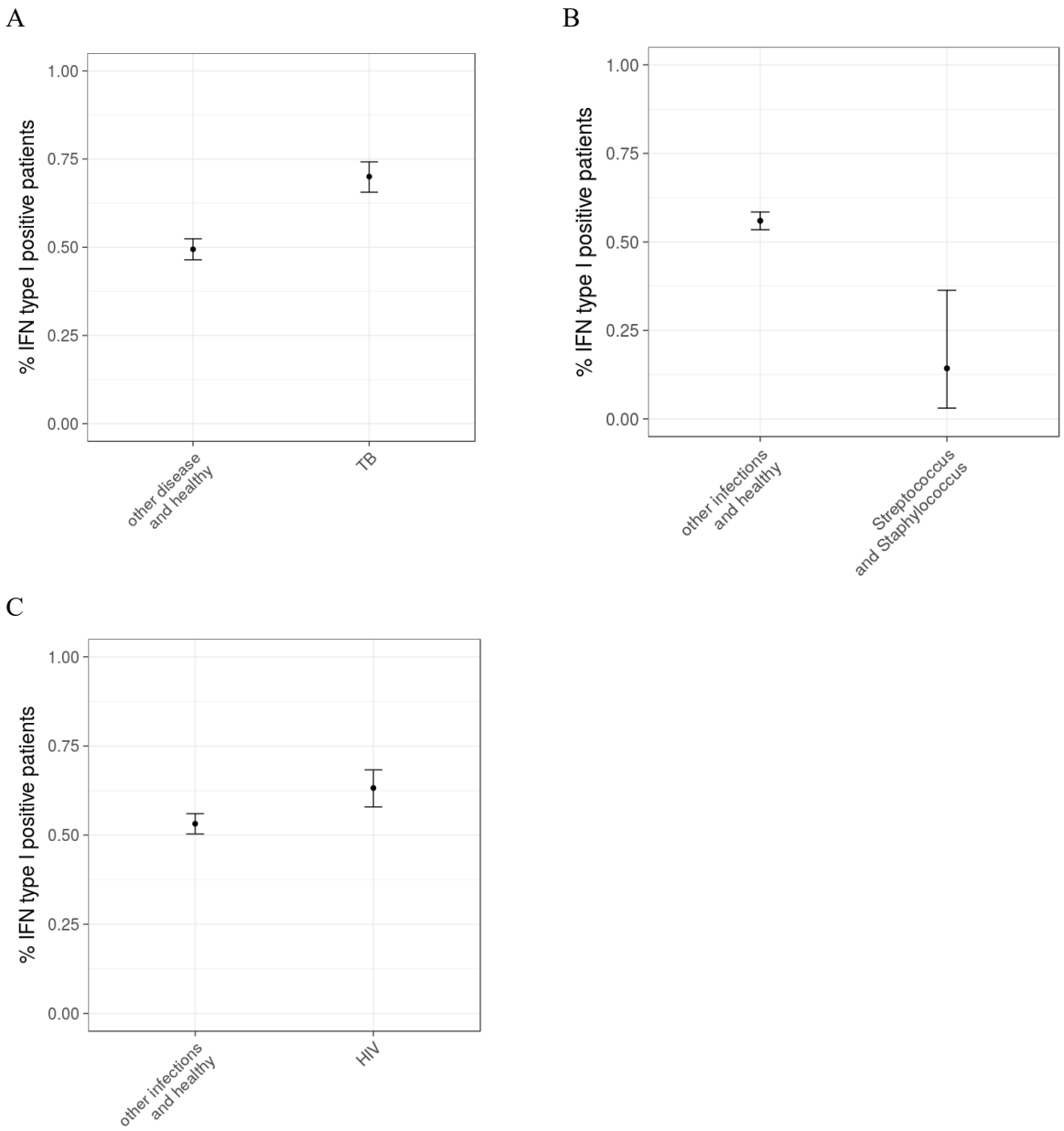


Figure 10 Percentage of IFN+ individuals in the MDS

(A) with and without TB, (B) with and without *Streptococcus sp.* and *Staphylococcus sp.* coinfection, (C) with and without HIV infection. The error bars indicate 95% confidence intervals. The 95% CI for Streptococcus and Staphylococcus coinfection are wider, because among 1574 donors in the MDS only 21 were coinfected with those two pathogens.

Clearly, active TB is the most significant predictor of the IFN+ status among the investigated factors, followed by HIV infection. The characteristics of the patients and different study details did not contribute significantly to the detection of IFN response. Nevertheless, there are TB patients who do not present IFN response.

I subsequently used PCA to further investigate the factors correlated with IFN enrichment. Even though clearly separated clusters were not present in the data, the IFN I+ and IFN I- samples clustered together which was best illustrated by the PCs 1 and 2 (Figure 11 A).

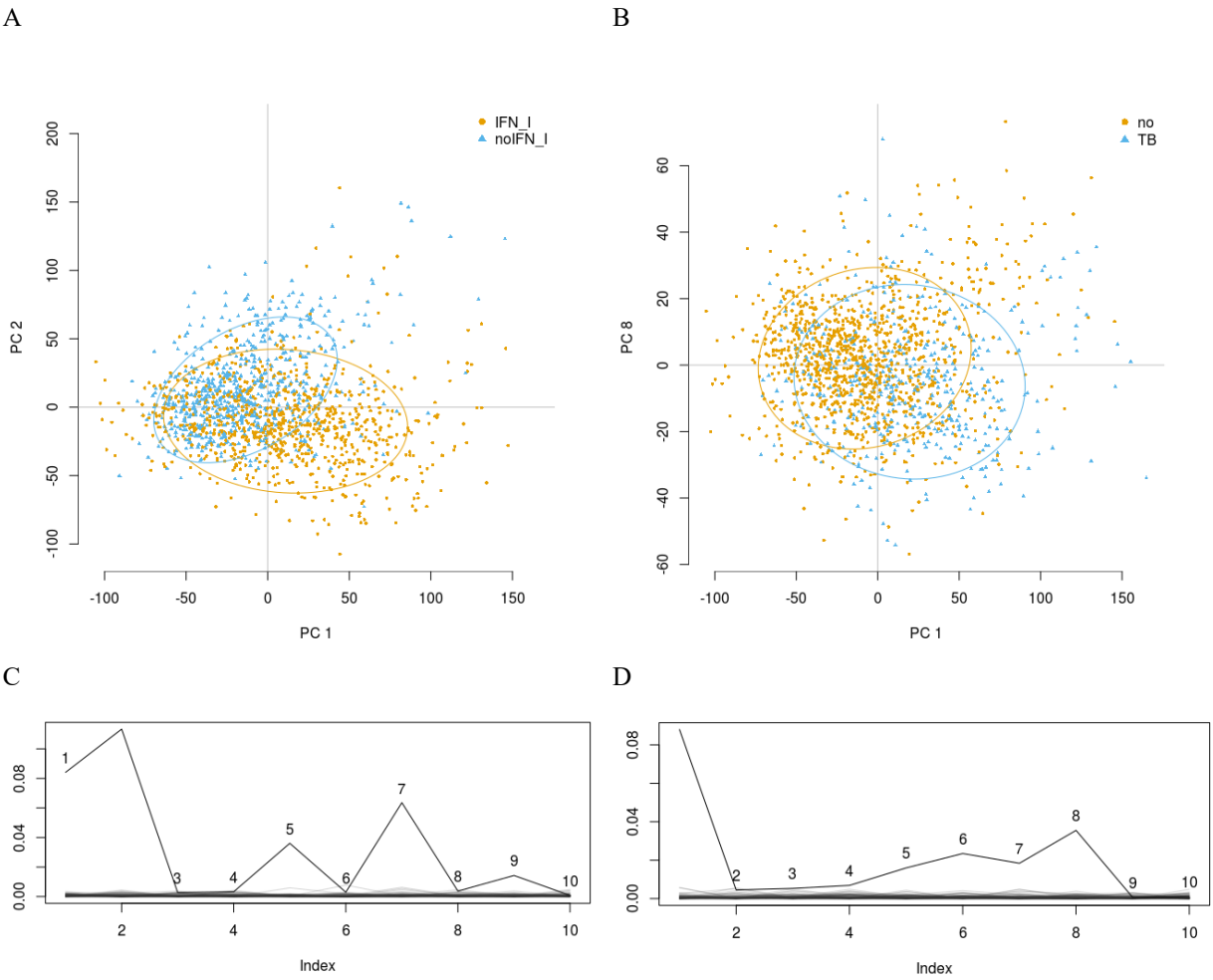


Figure 11 PCs of the matrix of gene expression in the training MDS

75% confidence interval ellipses for the groups are shown. (A) PC 1 and 2 colored by IFN status. (B) PC 1 and 8 colored by TB status. (C) Fraction of variance explained by IFN status as a predictor for each of the first 10 principal components of the gene expression matrix from MDS calculated using 100-times randomization. PC1 and PC2 explain the biggest fraction of the variance. (D) Fraction of variance explained by TB status as a predictor for each of the first 10 principal components of the gene expression matrix from MDS calculated using 100-times randomization. PC1 and PC8 explain the biggest fraction of the variance.

GSEA performed on the list of genes sorted according to their decreasing weights in PCA resulted in the lists of enriched modules which was dominated by enrichment in T- and NK-cell related modules for PC1 and in elements of innate immunity (including IFN response) as well as cell cycle and transcription in case of PC2. The complete list of modules enriched in PC1 and PC2 can be found in the Supplementary Table 5 and the enrichment of the modules with the p-value lower than 10^{-7} is visualized in the Figure 12 A, B.

The most significant variable according to the GLM influencing the IFN status - the presence of active TB was explained by the principal components (PCs) 1 and 8 (Figure 11). GSEA performed on the genes along those two PCs resulted in abundant list of the enriched modules typically found in TB patients (Supplementary Table 5, Figure 12 B, C), including IFN response, adaptive and innate immunity.

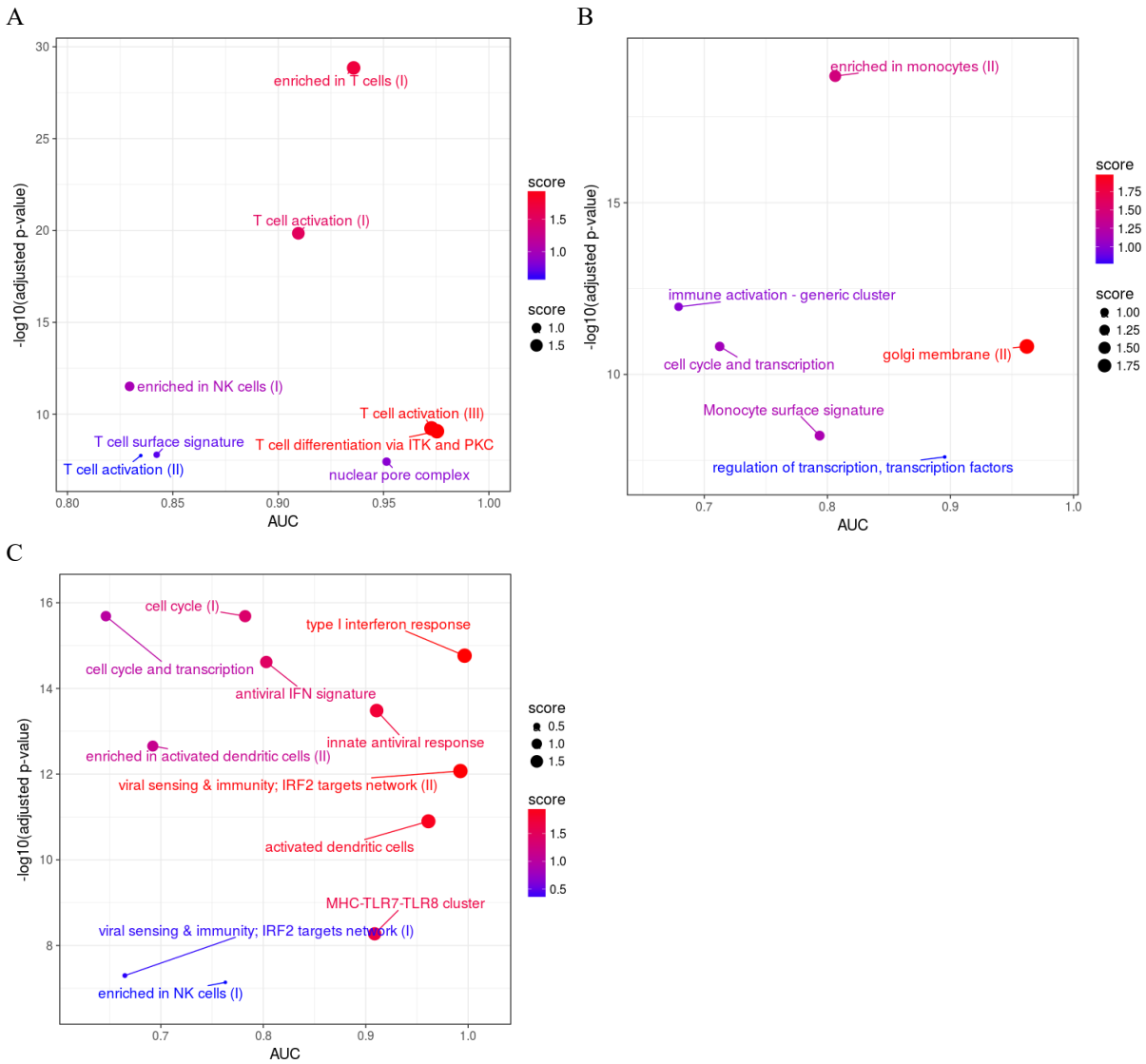


Figure 12 GSEA performed on the weighs of genes in PCs

(A) PC1, (B) PC2, (C) PC8 of the gene expression matrix from training MDS. For visualization purposes, only the modules enriched with p-value $<10^{-7}$ have been shown. The complete list of module enrichment can be found in the Supplementary Table 5. The score according to which the dots are colored is calculated by the *tmod* (Weiner & Domaszewska, 2016) package and is proportional to the rise in AUC and decrease of p-value of the module enrichment.

This showed that also unsupervised analysis identifies clusters of IFN- and IFN+ donors. The GSEA on the genes sorted by the weights in PCA indicated that T cell response is a major contributor to the differences seen between the IFN- and IFN+ patient clusters.

To investigate the dependence of IFN response on the study, ethnicity, residence, other diseases and HIV status among TB patients only, I created another logistic regression model using only the subset of data characterized by active TB. HIV- status and use of Illumina HumanHT-12 V3.0 platform were correlated with IFN+ status (Figure 13).

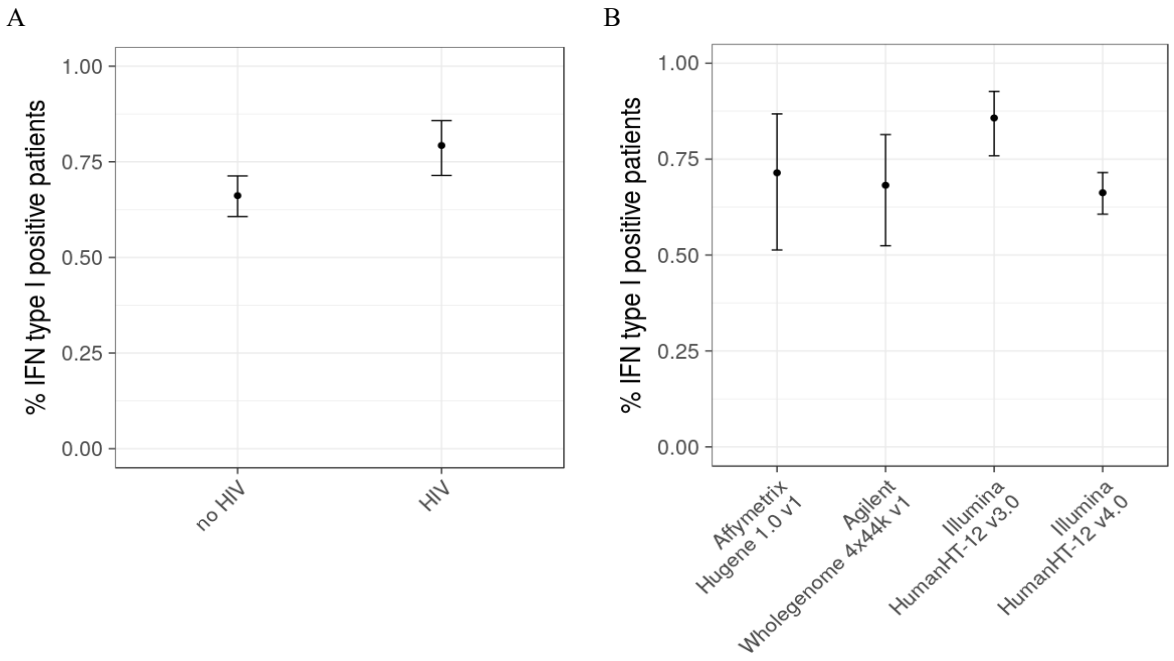


Figure 13 Percentage of IFN+ patients among TB patients from MDS

(A) with and without HIV coinfection, (B) investigated with different microarray platforms. The error bars represent 95% confidence intervals.

This means that even though normalization has been performed there are still platform-effects visible in the training MDS.

Calculation of the fraction of variance explained different factors using 100-times randomization revealed that in the case of data collected among TB patients the PCs 7 and 10 best illustrated the variance in the data influenced by platforms, PCs 2 and 6 best illustrated the variance explained by IFN status while the PCs 6 and 8 best illustrated the variance caused by HIV status (Figure 14).

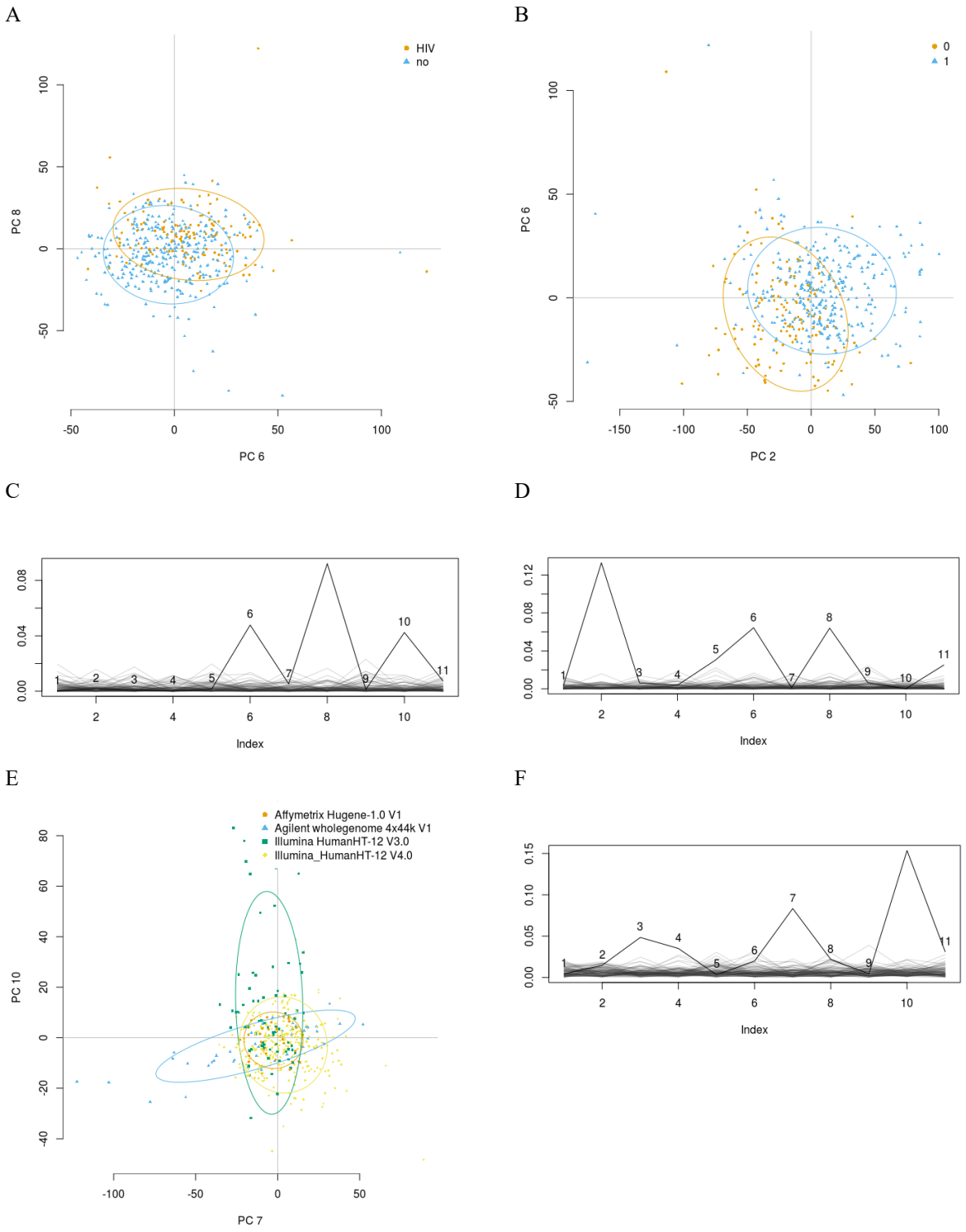


Figure 14 PCs of the matrix of gene expression of TB patients from the training MDS

75% confidence interval ellipses for the groups are shown. (A) PC 6 and 8 colored by HIV status. (B) PC 2 and 6 colored by IFN status. (C) Fraction of variance explained by HIV status as a predictor for each of the first 11 principal components of the gene expression matrix from TB patients from the training MDS calculated using 100-times randomization. PC6 and PC8 explain the biggest fraction of the variance. (D) Fraction of variance explained by IFN status as a predictor for each of the first 10 principal components of the gene expression matrix from TB patients from the training MDS calculated using 100-times randomization. PC2 and PC6 explain the biggest fraction of the variance. (E) PC 7 and 10 colored by used microarray platform. (F) Fraction of variance explained by the used microarray platform as a predictor for each of the first 11 principal components of the gene expression matrix from TB patients from the training MDS calculated using 100-times randomization. PC7 and PC10 explain the biggest fraction of the variance.

Again, a strong enrichment along the components 2 and 6 was detected, which differentiated between the IFN I+ and IFN I- patients (Figure 15 and Supplementary Table 6), with a dominating signature of T cells. Enrichment along the PC 8 explain together with PC 6 the biggest fraction of variance related to HIV status was related to cell cycle and IFN signaling. Enrichment in PC 7 presented only 5 significantly enriched modules while enrichment in PC 10 was non-significant, indicating that the platform effects did not convolute the biological effects.

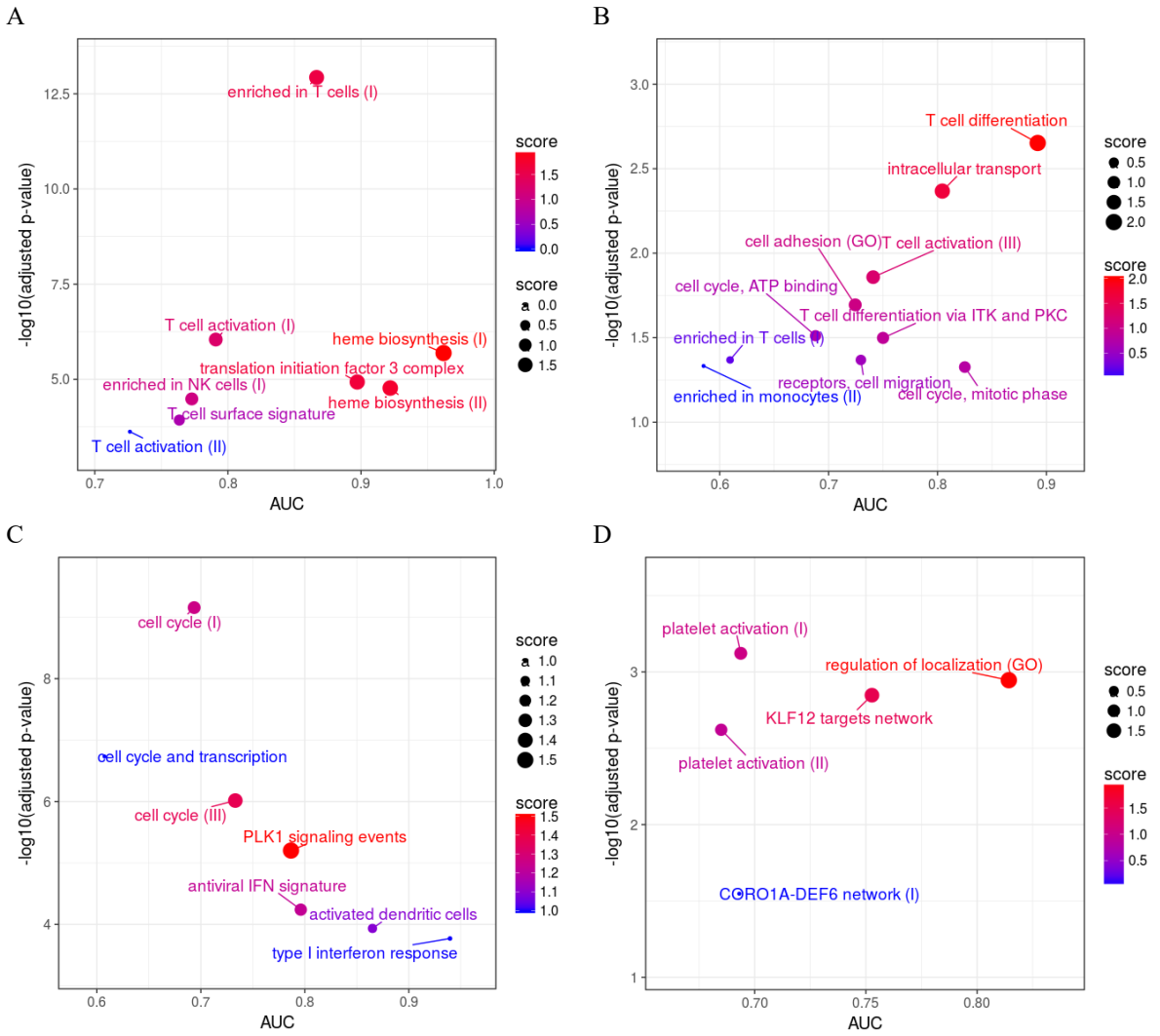


Figure 15 GSEA performed on the weighs of genes in PCs 2, 6, 7 and 8

(A) PC2, (B) PC6, (C) PC7, (D) PC8 of the gene expression matrix of TB patients from the training MDS. For visualization purposes, only the modules enriched with (A) p-value < $5 \cdot 10^{-4}$, (B) p-value <0.05, (C and D) p-value <10^{-3} have been shown. The complete list of module enrichment can be found in the Supplementary Table 6. The score according to which the dots are colored is calculated by the *tmod* (Weiner & Domaszewska, 2016) package and is proportional to the rise in AUC and decrease of p-value of the module enrichment.

Among TB patients the T cell and NK cell responses dominated the differences observed between IFN+ and IFN- patients. Furthermore, HIV infection was related to enrichment in IFN related modules. Analysis of the overall patient cohort as well as of TB patients indicated, that the IFN+ and IFN- status is strongly related to the presence of active TB but cannot be fully explained by any of the

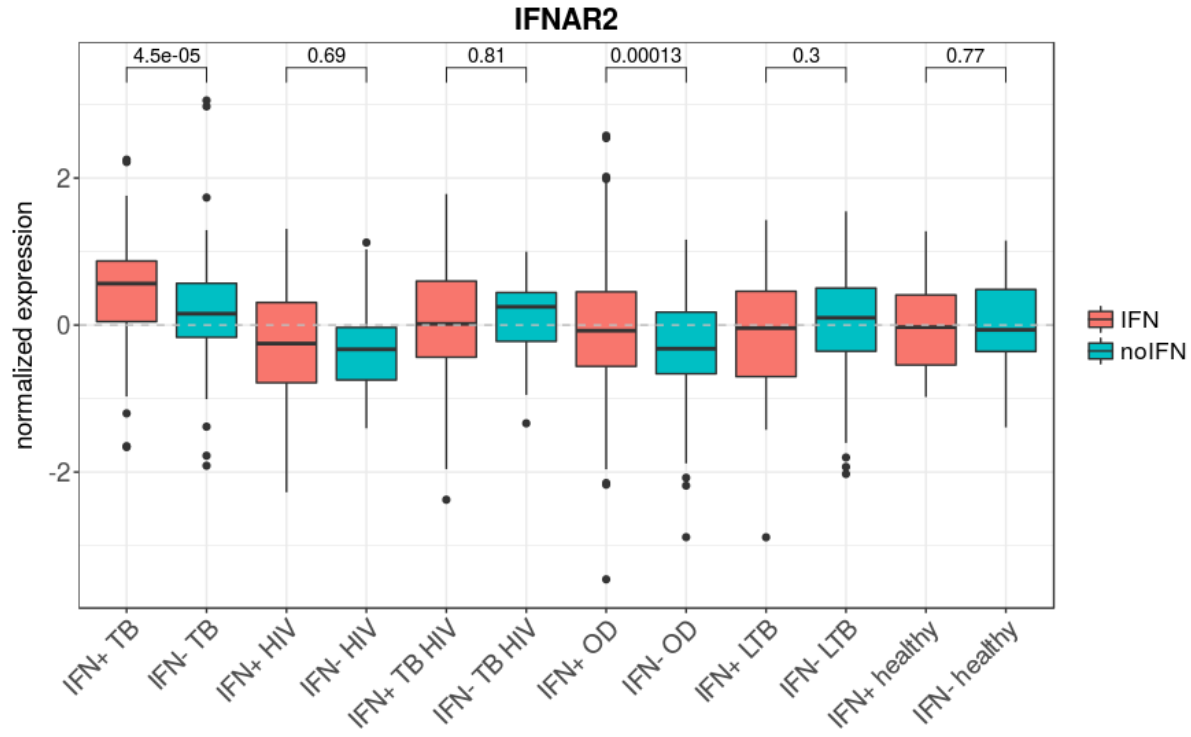
other investigated factors. Therefore, I continued to investigate the variability among TB patients that included the IFN status.

3.8. EXPRESSION OF INTERFERON-STIMULATED GENES IN THE BLOOD OF IFN+ AND IFN- PATIENTS

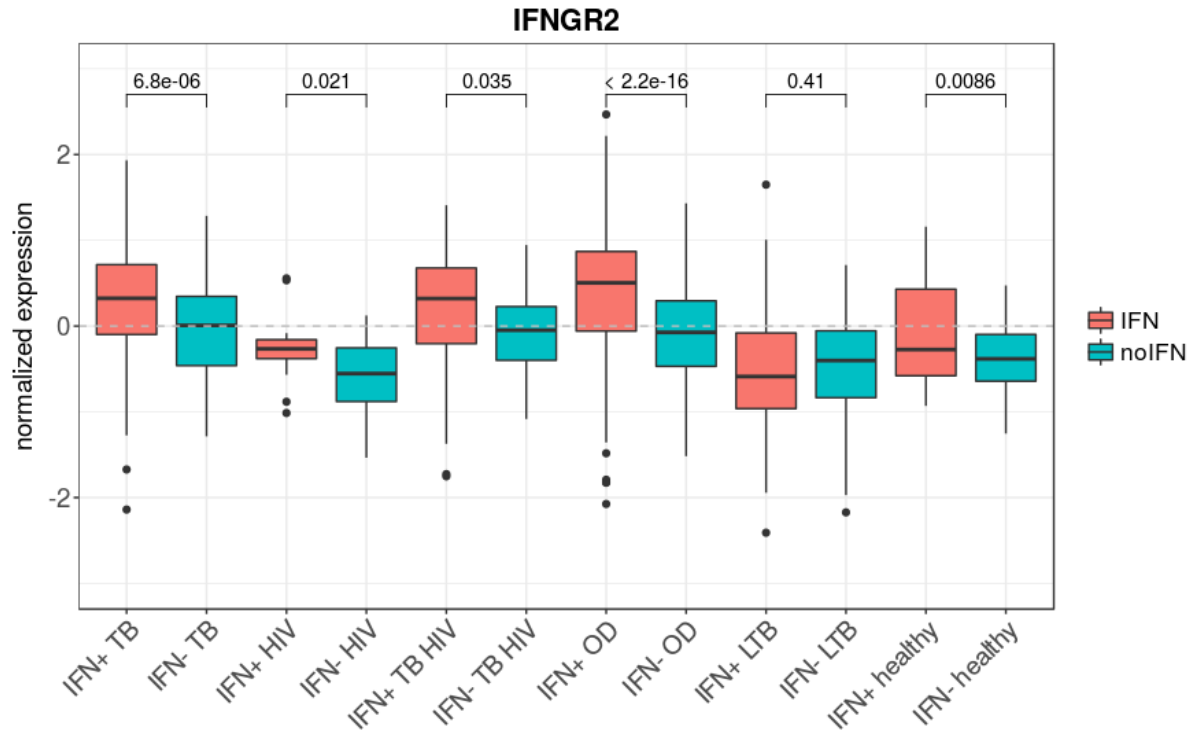
If the enrichment of IFN modules was a straightforward result of increased IFN-α, -β or -γ gene expression it would be possible to easily identify the correlation between the IFN status and the level of expression of the mentioned genes. Similarly, there could be such dependence of IFN receptor (IFNR) genes and the IFN status. Alternatively, if IFN and IFNR genes are not differentially expressed between the IFN+ and IFN- individuals it indicates, that the differences between IFN+ and IFN- patients should be related to the expression levels of the genes induced by IFNs. I investigated the expression of one IFN type I α, one IFN type I β and one IFN type II gene: IFNA2, IFNB1, IFNG, as well as their receptor genes: IFNAR2 and IFNGR2. I also compared the expression of three selected IFN-dependent genes: BATF2, CXCL10 and ANKRD22 in the IFN+ and IFN- groups of the six categories of donors: (i) TB positive, (ii) HIV positive, (iii) TB and HIV positive, (iv) OD, (v) LTB and (vi) healthy. The genes IFNGR2 and CXCL10 were present in the IFN gene modules used to identify IFN+ and IFN- patients, therefore the difference in the expression between IFN+ and IFN- patients was to be expected.

There were no significant differences in the expression of IFNA2, IFNB1 or IFNG genes between IFN+ and IFN- subgroups of any of the six categories of patients. Nevertheless, there were differences between the expression of IFNAR2 receptor genes between IFN+ and IFN- TB and OD patients (Figure 16). The expression of IFNGR2 gene was significantly different between IFN+ and IFN- subgroups in every category of donors except for LTBI. The expression of IFN-inducible genes BATF2, ANKRD22 and CXCL2 was significantly different between IFN+ and IFN- subgroups of every category of patients with exception of LTBI and HC for BATF2 and LTB for ANKRD22. The observed results suggest that the difference in the IFN+ and IFN- status of TB patients is not a result of increased expression of IFNA2, IFNB1 or IFNG genes. However, it can be related to the increased expression of IFN type I and type II receptor genes and it is depicted by increased levels of the transcripts of IFN inducible genes in the IFN+ individuals. It is remarkable that the initially used enrichment analysis indicated pathways which present significant differences between the IFN+ and IFN- patients which are now validated by comparing the expression of genes related to those pathways but were not included in the transcriptional modules. The significant differences in the expression of IFN-inducible genes between IFN+ and IFN- TB patients confirms that there are differences in the extent of transcriptional activation of IFN signaling pathways between the identified patient groups.

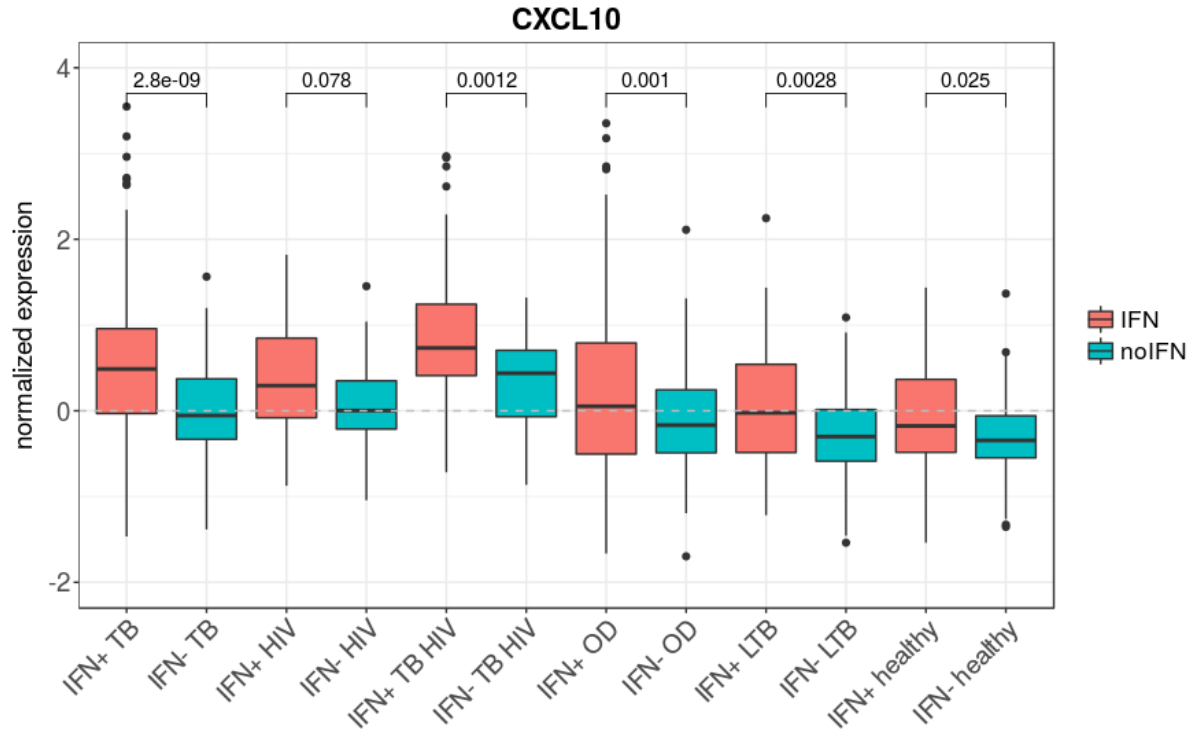
A


IFNAR2
4.5e-05
0.69
0.81
0.00013
0.3
0.77
normalized expression
2
0
-2
IFN+ TB
IFN- TB
IFN+ HIV
IFN- HIV
IFN+ TB HIV
IFN- TB HIV
IFN+ OD
IFN- OD
IFN+ LTB
IFN- LTB
IFN+ healthy
IFN- healthy
IFN
noIFN

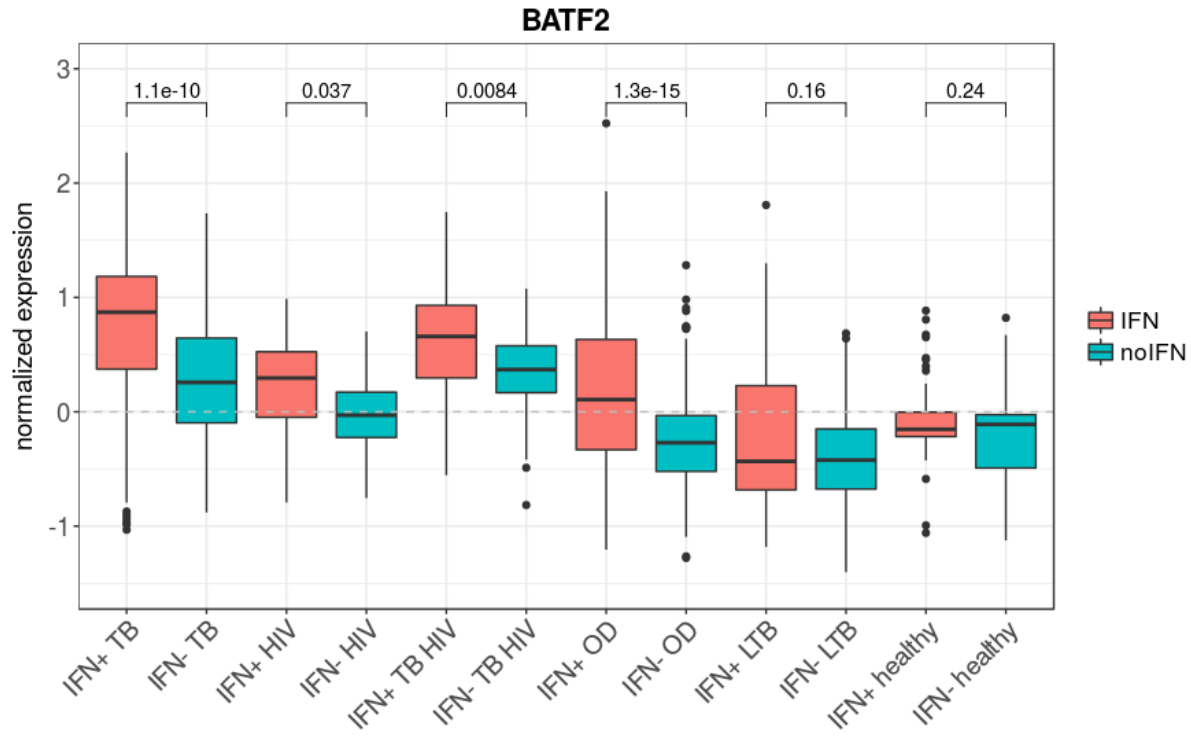
B


IFNGR2
6.8e-06
0.021
0.035
< 2.2e-16
0.41
0.0086
normalized expression
2
0
-2
IFN+ TB
IFN- TB
IFN+ HIV
IFN- HIV
IFN+ TB HIV
IFN- TB HIV
IFN+ OD
IFN- OD
IFN+ LTB
IFN- LTB
IFN+ healthy
IFN- healthy
IFN
noIFN

C


CXCL10
normalized expression
4
2
0
-2
2.8e-09
0.078
0.0012
0.001
0.0028
0.025
IFN+ TB
IFN- TB
IFN+ HIV
IFN- HIV
IFN+ TB HIV
IFN- TB HIV
IFN+ OD
IFN- OD
IFN+ LTB
IFN- LTB
IFN+ healthy
IFN- healthy
IFN
noIFN

D


BATF2
normalized expression
3
2
1
0
-1
1.1e-10
0.037
0.0084
1.3e-15
0.16
0.24
IFN+ TB
IFN- TB
IFN+ HIV
IFN- HIV
IFN+ TB HIV
IFN- TB HIV
IFN+ OD
IFN- OD
IFN+ LTB
IFN- LTB
IFN+ healthy
IFN- healthy
IFN
noIFN

E

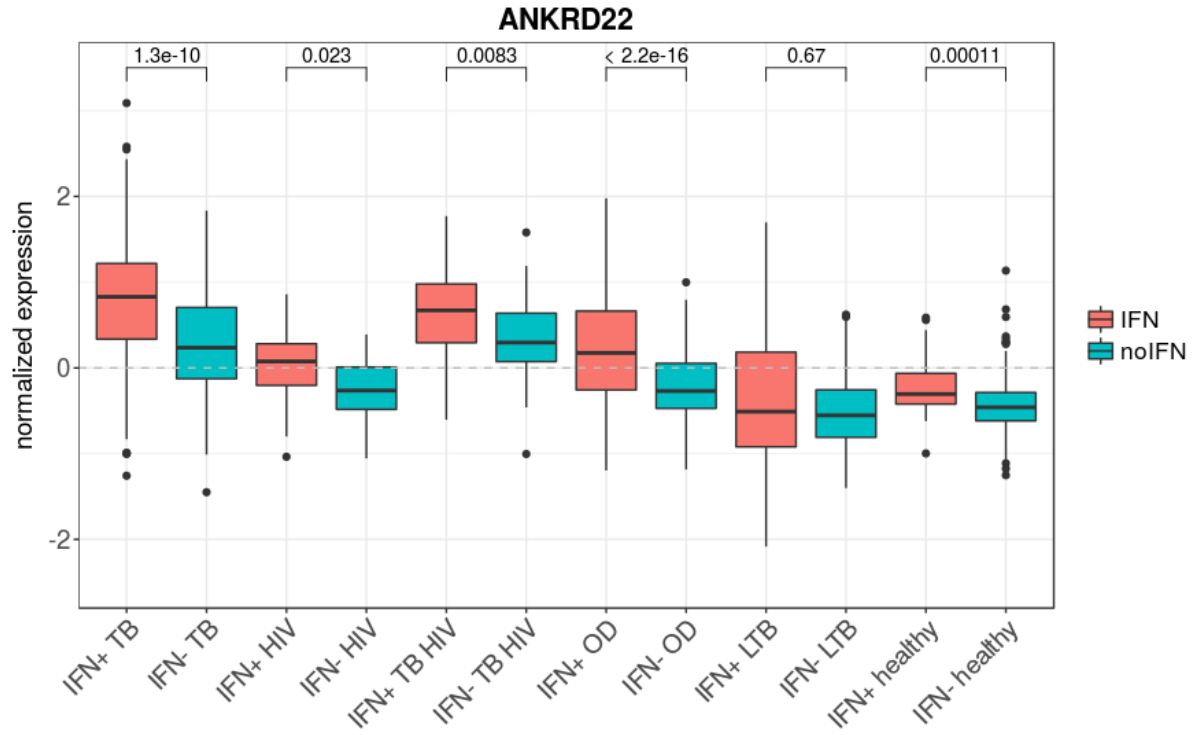


Figure 16 Expression of IFN type I and type II related genes in the IFN+ and IFN- subgroups of TB positive, HIV positive, HIV and TB positive, OD patients, LTB and HCs

The differences in the gene expression level of those genes is most significant between IFN+ and IFN- TB and OD patients. (A) IFNAR2, (B) IFNGR2, (C) CXCL10, (D) BATF2, (E) ANKRD22 gene expression is shown.

3.9. THE EXPRESSION OF SEVERAL IMPORTANT GENES FOR TB IS MARKEDLY DIFFERENT BETWEEN IFN+ AND IFN- PATIENTS

The division into IFN+ and IFN- TB patients could reveal other non-IFN related genes or gene modules that exist within these groups. To investigate if there are significant discordant genes between IFN- and IFN+ TB patients apart from the IFN-inducible genes, IFN genes and IFNR genes I calculated differential expression between IFN+ TB patients and healthy and IFN- TB patients and healthy. The lists of differentially expressed genes are available on the website: http://bioinfo.mpiib-berlin.mpg.de/TBprofiles/. I calculated disco.score for the pairs of corresponding genes which is presented in the Figure 17. Disco.score analysis presents a different approach to comparing gene expression than the differential expression analysis. It results in an ordered list of genes sorted by similarity of their expression regulation in two groups. Therefore, it can indicate genes or gene modules which are characterized by opposite expression regulation (i.e. overexpression vs underexpression) as well as the genes with marked differences in the expression between two groups even though the expression is concordant (i.e. underexpression in both groups or overexpression in both groups).

There were no significantly discordant genes between the IFN+ and IFN- TB patients. However, the scale of gene regulation was markedly different in the two patient groups and several genes, including CD273, CD274, C1QC, PAR2, which were upregulated in both groups were characterized by around twice as high log_2FC in IFN + as in IFN- patients. All of those genes have been previously reported to play a role in TB which I further discuss in the Chapter 5.

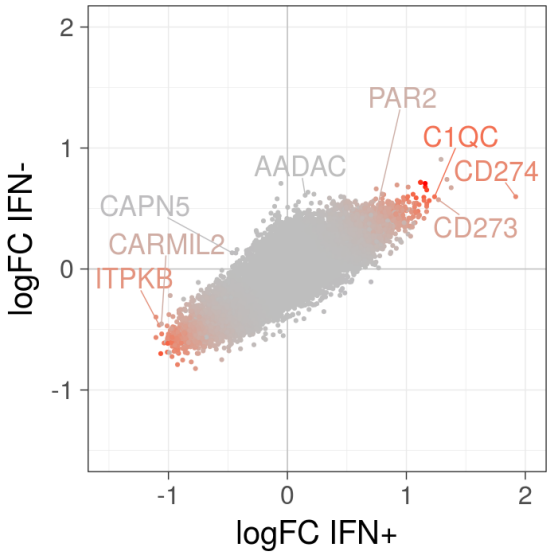


Figure 17 Concordant and discordant genes between the IFN+ and IFN- TB patients

Increasing intensity of the red color indicates increase in disco.score and illustrates higher degree of similarity between gene expression in the two patient cohorts. Increasing intensity of the blue color which is not observed in this figure would indicates decrease in negative disco.score and a higher degree of dissimilarity in gene expression between two patient cohorts.

3.10. CYTOKINE LEVELS IN BLOOD CORRESPOND TO THE IFN I+/IFN I- STATUS

I have shown that the GSEA-based division into IFN+ and IFN- patients corresponds with the levels of transcripts of IFN-inducible genes in blood of the IFN+ and IFN- TB patients. However, to show that this division reflects what is actually happening in blood on the level of IFN-inducible cytokines I needed to use a dataset where transcriptomic data would be accompanied by the measurements of cytokines in blood. Correlation of the IFN+ status with the increased level of IFN inducible cytokines in blood would be a direct proof of the functional consequence of the suggested GSEA-based division into IFN+ and IFN- patients.

The expression of CXCL2 and CCL2 chemokines is triggered by IFN signaling. To compare the levels of those cytokines in WB with the IFN+/IFN- status of the patients, I used data from a study cohort where healthy volunteers were vaccinated with IFN response inducing influenza vaccination – FLUAD™. The datasets contained (i) gene expression measured by microarrays and (ii) cytokine absolute concentrations in WB from 114 volunteers, measured before the FLUAD™ influenza vaccination and 1 day after the vaccination. To identify the IFN+ and IFN- individuals I used z-scores

calculated based on normalized gene expression values as described before, using average concentrations of the respective cytokines in the samples before vaccination as reference values.

Based on the GSEA using IFN modules the samples collected before the vaccination were classified as IFN I-, whereas majority of the samples collected the day after the vaccination as IFN I+ (Figure 18). Four out of six placebo control samples were classified as IFN I-. The mean values of both CXCL10 and CCL2 concentration for study participants identified as IFN I+ were higher than the values for IFN- participants. Moreover, ROC curve for the IFN status based on the cytokine level as binary predictor showed that the concentration of CXCL10 in blood is a specific and sensitive classifier for IFN I+ status and CCL2 presents rather low sensitivity and specificity as a classifier for the IFN I+ status (Figure 19).

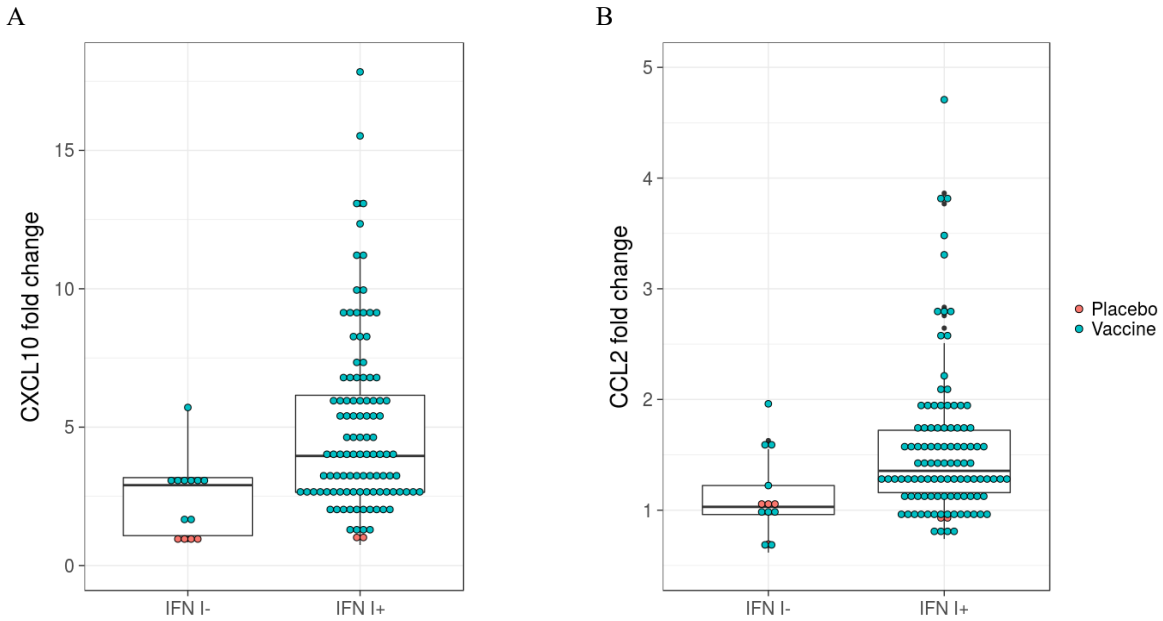


Figure 18 Fold changes of WB cytokine levels of volunteers vaccinated with FLUAD vaccine in day 1 after vaccination compared to the vaccination day

Presented are levels of IFN inducible cytokines: CXCL10 (A) and CCL2 (B). The data points marked red are derived from Placebo-injected individuals.

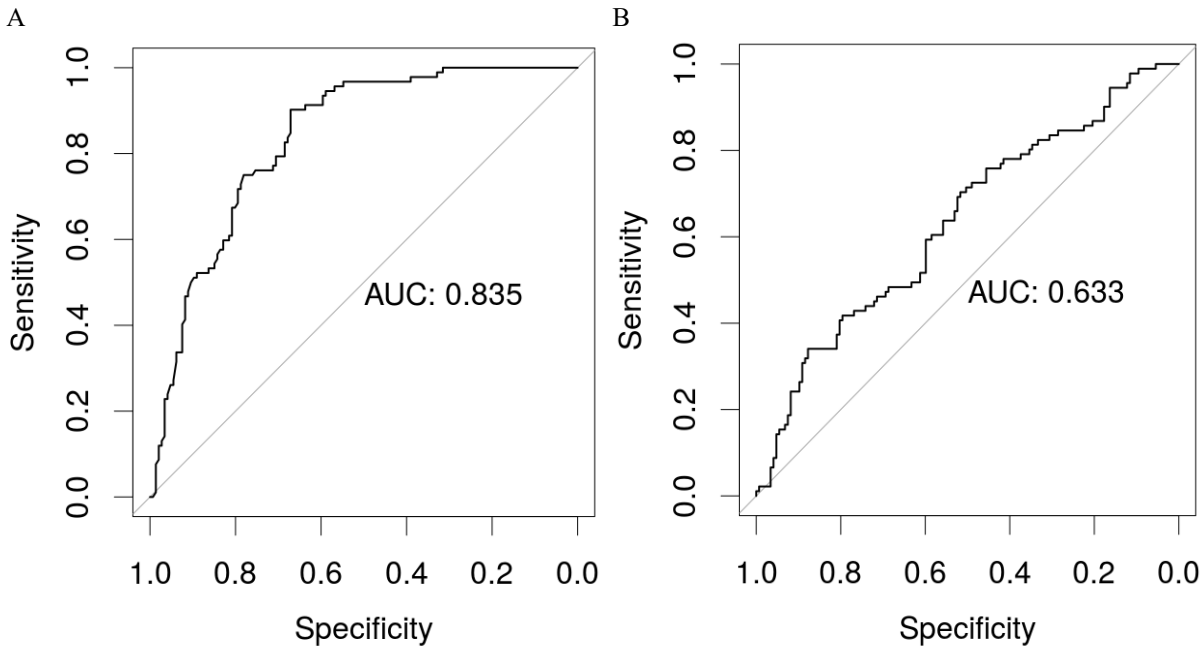


Figure 19 ROC curves characterizing the sensitivity and specificity of CXCL10 and CCL2 as binary predictors of IFN status

The level of IFN inducible cytokines CXCL2 **(A)** and CCL2 **(B)** can be used as binary predictors of IFN status.

Therefore, I showed that the division into IFN+ and IFN- individuals based on the gene expression profiles is reflected by the levels of IFN inducible CXCL10 and CCL2 cytokines in blood.

3.11. CORRELATION BETWEEN INTERFERON STATUS AND THE DISEASE SEVERITY

In the study by Berry et al. (2010) 80 randomly chosen study participants underwent lung X-Ray investigation and the results of 72 of the lung images were classified by three independent physicians blinded to microarray data and clinical diagnosis to one of the four categories: (i) no disease, (ii) minimal disease, (iii) moderate disease, (iv) advanced disease. I defined IFN I+/IFN I- status for the participants of the study by Berry et al. and compared it with the X-Ray based disease classification. 26% of the donors classified into "no disease" category presented IFN I+ status (Figure 20). Among the patients with "minimal disease" 57% were IFN I+, and in both categories "moderate" and "advanced" disease over 90% of the patients were IFN I+.

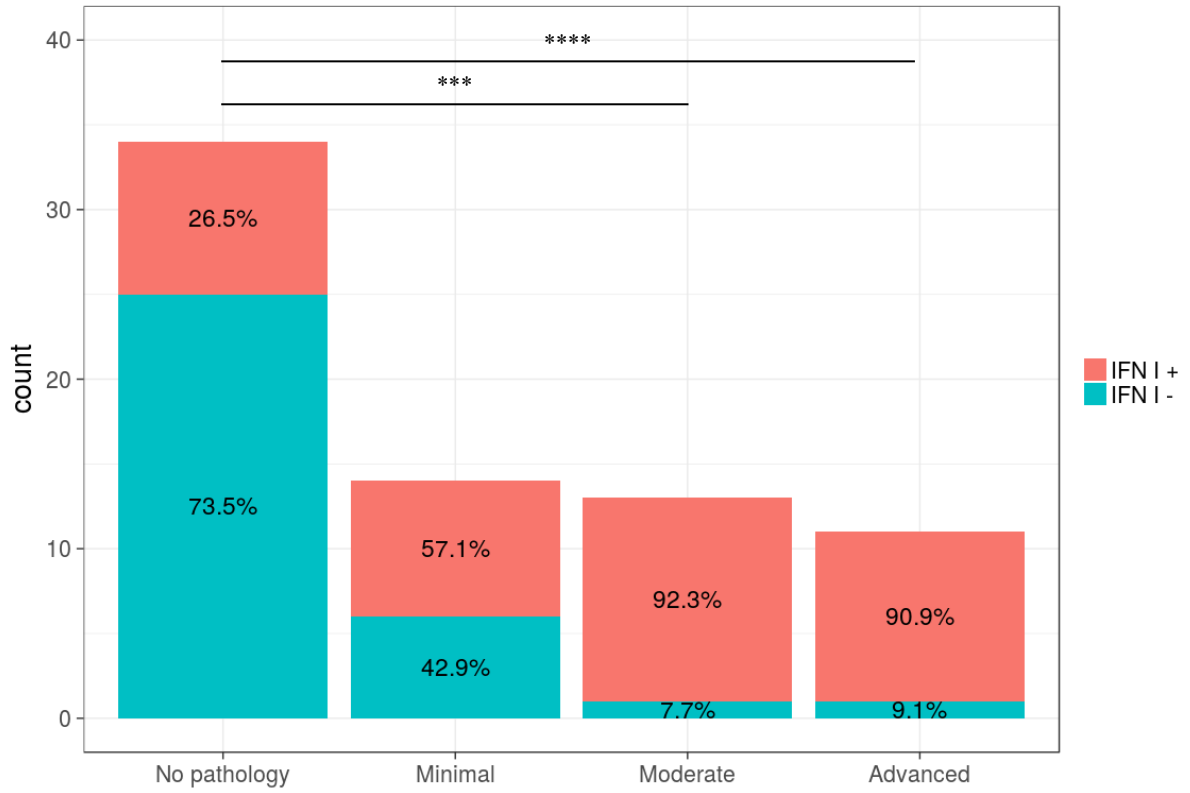


Figure 20. IFN status of the patients with varying levels of pathology in lungs

Almost all patients with moderate and advanced pathology present IFN+ status. The proportion of the IFN+ and IFN- patients was significantly different between the patients with no pathology compared to the patients with moderate ($p = 1 \cdot 10^{-3}$ in pairwise comparisons using Fisher's exact test for count data with Bonferroni correction) and with advanced pathology ($p = 4 \cdot 10^{-4}$).

This indicated that the activation of IFN signaling pathways on gene expression level corresponds with advanced pathology in the lungs of TB patients.

3.12. RANDOM FOREST CLASSIFICATION

I used RF models to see how IFN status influences the ability of the models to classify TB patients. The preliminary RF models were trained using 10-fold cross-validation to distinguish between (i) IFN+ TB patients and healthy individuals, (ii) IFN- TB patients and healthy individuals, (iii) IFN+ TB patients and all non-TB individuals, (iv) IFN- TB patients and all non-TB individuals, (v) IFN+ patients and OD, and (vi) IFN- TB patients and OD (Table 8). Six models were created using the whole set of genes measured in the MDS and six models with exclusion of IFN I genes to investigate how the model's performance changes when the IFN type I genes are excluded; specifically, whether the performance of the model based on the IFN+ individuals is more similar to the performance of the models based on IFN- individuals. The models were tested for their sensitivity and specificity in discriminating (i) TB patients from non-TB donors, (ii) IFN+ TB patients from healthy donors, (iii) IFN- TB patients from healthy donors, (iv) IFN+ TB patients from patients with OD, (v) IFN- TB patients from patients with OD.

Table 8. Characteristics of the preliminary RF models

Six RF models were built using all the genes in MDS (Models 1, 2, 5, 6, 9, 10) and 6 RF models were built with exclusion of the genes present in IFN type I module set (Supplementary Table 2). The models 1 - 4 were built using TB and healthy individuals, models 5 – 8 using TB and all non-TB individuals and models 9 – 12 using TB and OD patients. The models were built including either IFN+ (models 1, 3, 5, 7, 9, 11) or IFN- TB patients (models 2, 4, 6, 8, 10, 12).

	All genes		No IFN I genes	
	TB IFN +	**TB IFN -**	**TB IFN +**	**TB IFN -**
healthy	Model 1	Model 2	Model 3	Model 4
non TB	Model 5	Model 6	Model 7	Model 8
OD	Model 9	Model 10	Model 11	Model 12

The models trained on the IFN+ TB patients showed high sensitivity and specificity in differentiating between IFN+ TB patients and healthy, AUC = 0.96 (95% CI = 0.95-0.97), but lower sensitivity and specificity in distinguishing IFN- patients from healthy donors, AUC = 0.81 (95% CI = 0.77-0.86; Figure 21 A). Their ability to distinguish IFN- patients from OD patients was slightly better than random, AUC = 0.57 (95% CI = 0.52-0.62). Interestingly, the models trained on the IFN- patients and healthy individuals also showed highest sensitivity and specificity in distinguishing between IFN+ and healthy, AUC = 0.95 (95% CI = 0.94-0.97) and lower in distinguishing IFN- patients from healthy, AUC = 0.87 (95% CI = 0.83-0.91). They were not capable of distinguishing IFN- TB patients from patients with OD, AUC = 0.47 (95% CI = 0.41-0.52) and the overall sensitivity and specificity in distinguishing TB from non TB similar as in the case of the model based on the IFN + patients, AUC = 0.77 (95% CI = 0.75-0.80; Figure 21 B). Both models showed significant differences when identifying

TB patients among different groups of patients (healthy, non-TB and OD). This can be understood as a flaw of the model, since an ideal model should identify any TB patients among healthy as well as OD patients.

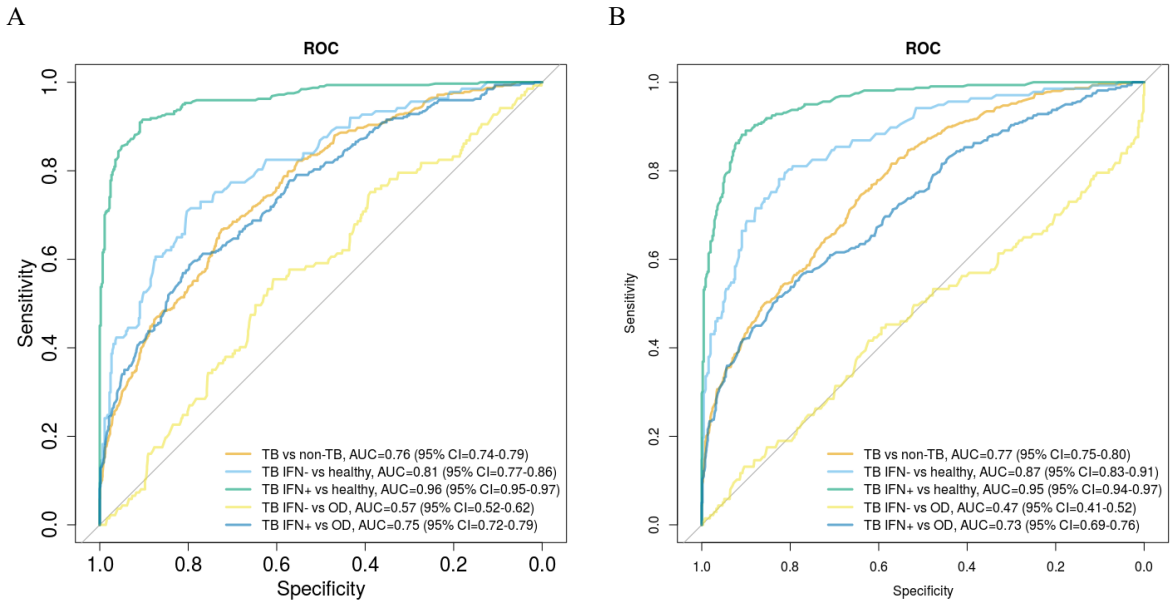


Figure 21 Results of testing the RF models 1 and 2 using k-fold cross validation

Model 1 (A) was trained to differentiate between IFN+ TB patients and healthy and tested for differentiating both subgroups of TB patients (TB IFN+, TB IFN-) from healthy (ROCs described as "vs healthy" in the figure legend) and from OD (ROCs described as "vs OD" in the figure legend) donors. Model 2 (B) was trained to differentiate between IFN- TB patients and healthy and tested for differentiating both subgroups of TB patients (TB IFN+, TB IFN-) from healthy and from OD donors. The overall performance of the models is presented by the ROCs described as "TB vs non-TB".

To investigate the influence of the presence of IFN I genes in the model I repeated the analysis excluding the genes present in IFN I module set from the MDS thus creating the models 3 and 4 (Table 8). The resulting ROC curves were nearly identical to the ones in the Model 1 and Model 2, which indicated that the IFN genes did not significantly improve the diagnostic capability of the models (Figure 22).

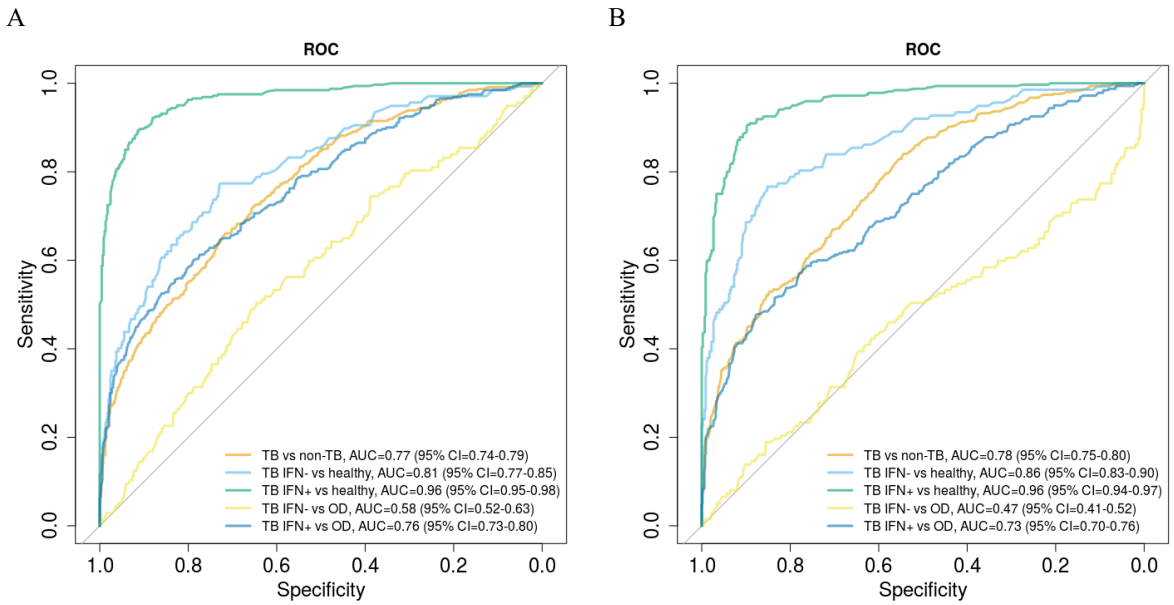


Figure 22 Results of testing the models 3 and 4 using k-fold cross validation

Models 3 (A) and 4 (B) were created with exclusion of genes present in IFN modules. Model 3 was trained to differentiate between IFN+ TB patients and healthy and tested for differentiating both subgroups of TB patients (TB IFN+, TB IFN-) from healthy (ROCs described as "vs healthy" in the figure legend) and from OD donors (ROCs described as "vs OD" in the figure legend). Model 4 was trained to differentiate between IFN- TB patients and healthy and tested for differentiating both subgroups of TB patients (TB IFN+, TB IFN-) from healthy and from all non-TB donors. The overall performance of the models is presented by the ROCs described as "TB vs non-TB".

Models 5, 6 7 and 8 were created analogously, but trained to distinguish between the IFN I+ or IFN I- TB patients and all non TB individuals present in the study, including LTBI, healthy and patients with other diseases. In this case the models trained on the IFN I- presented much more stable behavior (Figure 23). The sensitivity and specificity of distinguishing IFN- TB patients from the healthy donors or OD patients was increased in comparison to the models trained on IFN+ patients (AUC TB IFN- vs healthy = 0.81 (95% CI = 0.77-0.85), AUC TB IFN- vs OD = 0.80 (95% CI = 0.75 – 0.84); Figure 23 A, B), while the sensitivity and specificity of distinguishing IFN+ TB patients from the healthy donors or OD patients decreased in comparison to the model trained on IFN+ TB patients and non-TB donors (AUC TB IFN+ vs healthy = 0.78 (95% CI = 0.74-0.81), AUC TB IFN+ vs OD = 0.76 (95% CI = 0.72 – 0.79)). Overall the model performance was better than of the comparable models trained against only healthy donors (AUC TB IFN+ vs non TB = 0.82 (95% CI = 0.80 – 0.85), AUC TB IFN- vs non TB = 0.77 (95% CI = 0.75 – 0.80)). The exclusion of IFN genes did not influence the model performance (Figure 23 C, D).

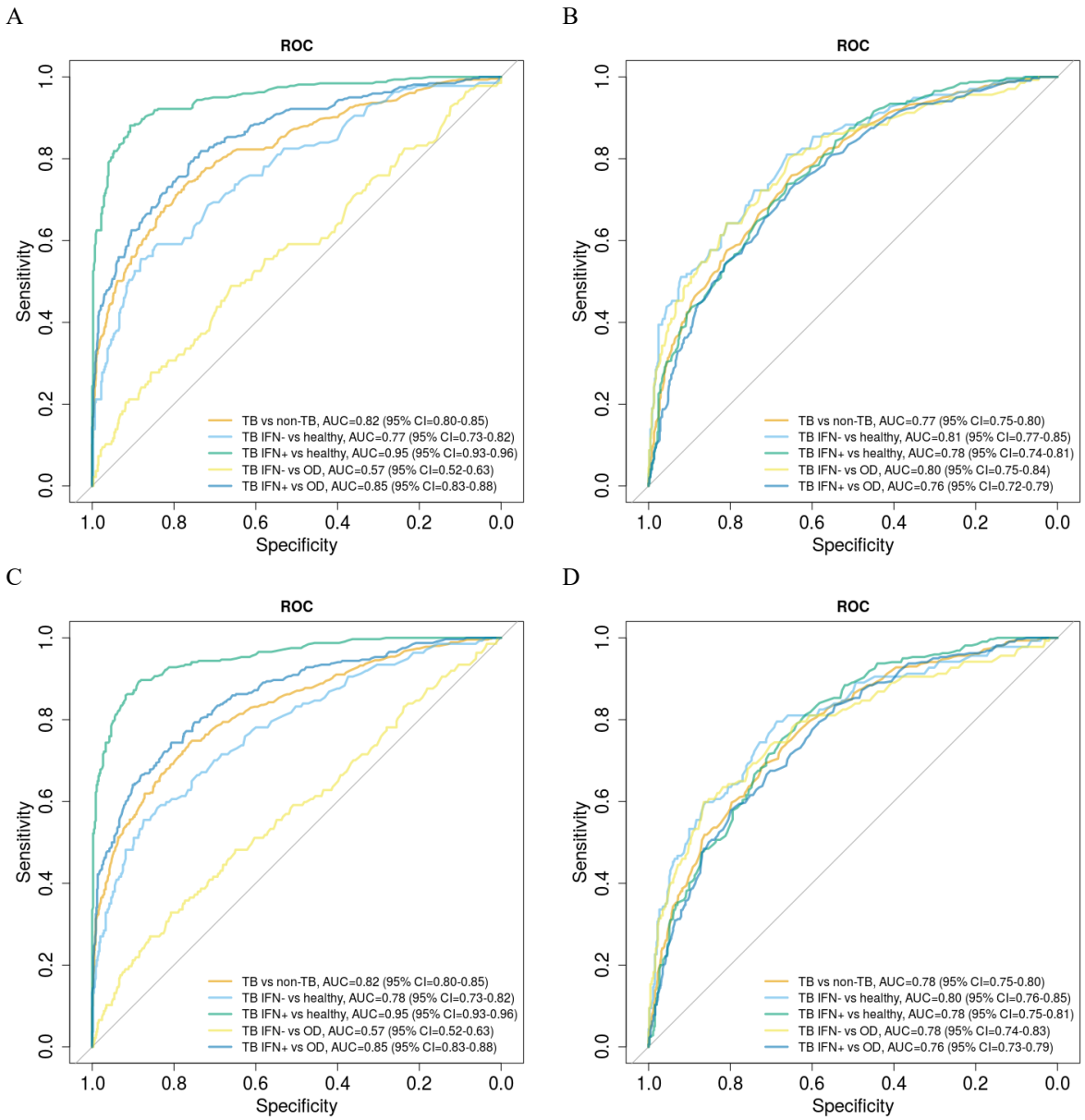


Figure 23. Results of testing the models 5, 6, 7 and 8 using k-fold cross validation

Models 5 (A) and 6 (B) were created using all genes present in MDS while models 7 (C) and 8 (D) were created with exclusion of genes present in IFN modules. Models 5 and 7 were trained to differentiate between IFN+ TB patients and non-TB and tested for differentiating both subgroups of TB patients (TB IFN+, TB IFN-) from healthy (ROCs described as "vs healthy" in the figure legend) and from OD (ROCs described as "vs OD" in the figure legend) donors. Models 6 and 8 were trained to differentiate between IFN- TB patients and all non-TB and tested for differentiating both subgroups of TB patients (TB IFN+, TB IFN-) from healthy and from OD donors. The overall performance of the models is presented by the ROCs described as "TB vs non-TB".

The last four models were created analogically but trained to distinguish between IFN I+ or IFN I- TB patients from patients with OD. In this case, the performance of the model trained on the IFN I- TB patients was characterized by significantly lower overall sensitivity and specificity than the models trained on the IFN I+ TB patients (AUC TB vs non TB trained on IFN- TB and OD = 0.63 (95% CI = 0.60-0.66), AUC TB vs non TB trained on IFN+ TB and OD = 0.82 (95% CI = 0.80-0.85), Figure 24 A, B); interestingly, those parameters were better in Model 9 than in the Model 10 for all patient

groups except for IFN I- TB patients. Again, exclusion of the IFN genes did not significantly influence the model performance (Figure 24 C, D).

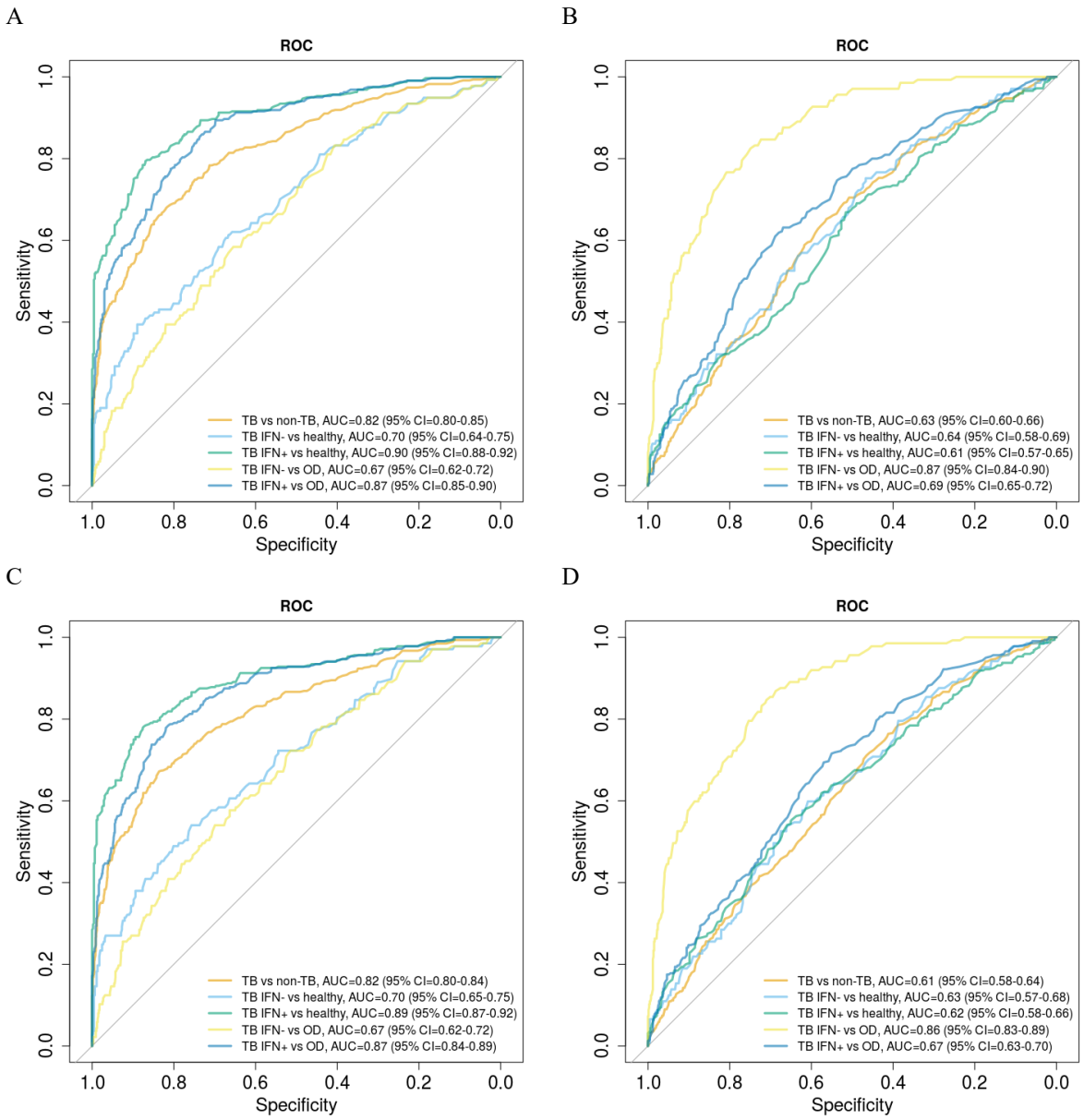


Figure 24. Results of testing of the models 9, 10, 11 and 12 using k-fold cross validation

Models 9 (A) and 10 (B) were created using all genes present in MDS while models 11 (C) and 12 (D) were created with exclusion of genes present in IFN modules. Models 9 and 11 were trained to differentiate between IFN+ TB patients and OD and tested for differentiating both subgroups of TB patients (TB IFN+, TB IFN-) from healthy (ROCs described as "vs healthy" in the figure legend) and from OD (ROCs described as "vs OD" in the figure legend) donors. Models 10 and 12 were trained to differentiate between IFN- TB patients and OD and tested for differentiating both subgroups of TB patients (TB IFN+, TB IFN-) from healthy and OD. The overall performance of the models is presented by the ROCs described as "TB vs non-TB".

In summary, the most stable were the models trained on IFN- TB patients and non-TB individuals (Model 6 and 8). They were characterized by the overall AUC of 0.77 and 0.78, correspondingly. I summarize the results of all created models in the Figure 25.

The modes 6 and 8 were not only most robust in a sense of being trained to identify TB patients among all non-TB individuals, but also characterized by the best cumulative AUC. The model trained on the IFN- and non-TB donors was characterized by the highest stability.

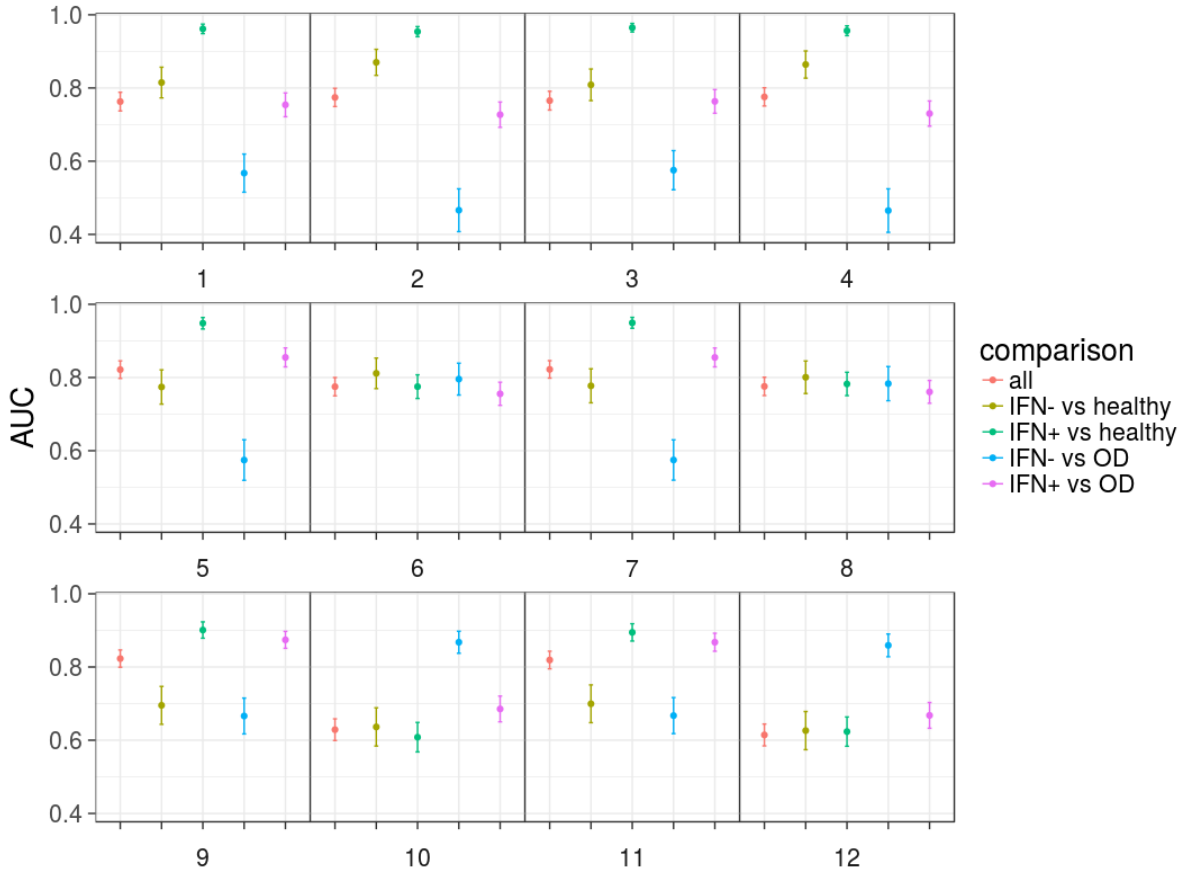


Figure 25 Summary of the performance of the created RF models

The preliminary RF models were trained using 10-fold cross-validation to distinguish between (i) IFN+ TB patients and healthy individuals (Panel 1 and 3), (ii) IFN- TB patients and healthy individuals (Panel 2 and 4), (iii) IFN+ TB patients and all non-TB individuals (Panel 5 and 7), (iv) IFN- TB patients and all non-TB individuals (Panel 6 and 8), (v) IFN+ patients and OD (Panel 9 and 11), and (vi) IFN- TB patients and OD (Panel 10 and 12). Six models were created using the whole set of genes measured in the MDS (Panels 1, 2, 5, 6, 9, 10) and six models with exclusion of IFN I genes (Panels 3, 4, 7, 8, 11, 12). For each model the AUC with 95% confidence intervals is shown for identification of: TB patients among all non-TB patients in a given model (red); IFN- TB patients among healthy patients in a given model (olive green); IFN+ TB patients among healthy patients in a given model (green); IFN- TB patients among OD patients in a given model (blue) and IFN+ TB patients among OD patients in a given model (pink). The most stable results are given by the models 6 and 8. Models 3 and 4, 7 and 8, and 11 and 12 are almost identical as models 1 and 2, 5 and 6 and 9 and 10, correspondingly. This means that the exclusion of genes present in IFN modules does not significantly influence the performance of the models.

3.13. BIOSIGNATURES OF THE IFN + AND IFN - TB PATIENTS

The biosignature of TB patients should be sufficiently specific to distinguish them not only from healthy blood donors but also from patients with OD. At the same time, the models trained on the IFN- and IFN+ TB patients and non-TB individuals were characterized by the best cumulative AUC and IFN- and non-TB based model by the best stability. Therefore in order to identify biosignatures of IFN+ and IFN- TB patients I trained the RF models using (i) IFN+ TB patients and non-TB individuals containing healthy, LTBI as well as OD patients and (ii) IFN- TB patients and non-TB individuals containing healthy, LTBI as well as OD patients, corresponding to the model 5 and model 6 (Table 8). The biosignature should ideally consist of a small number of genes which still retains high sensitivity and specificity. To choose the appropriate biosignature size I first investigated how the AUC of ROC curves presenting sensitivity and specificity of classification of the individuals as TB or non-TB individuals depend on the signature size. Therefore I divided the training set into 10 folds and using 9 of them I created RF models for (i) IFN+ TB patients and non-TB and (ii) IFN- TB patients and non-TB individuals and derived biosignatures of 6 different sizes: 5 genes, 7 genes, 10 genes, 20 genes, 50 genes and 200 genes. In the next step, I used the remaining fold as a test set to estimate the performance of the derived biosignatures. The 20 gene biosignature showed optimal performance for classification of TB and non-TB individuals based on the model trained on IFN+ TB patients and non-TB individuals, achieving AUC of 0.83 in the test set. For the model built on IFN- TB patients and non-TB individuals the optimal signature size was 50 genes (Figure 26).

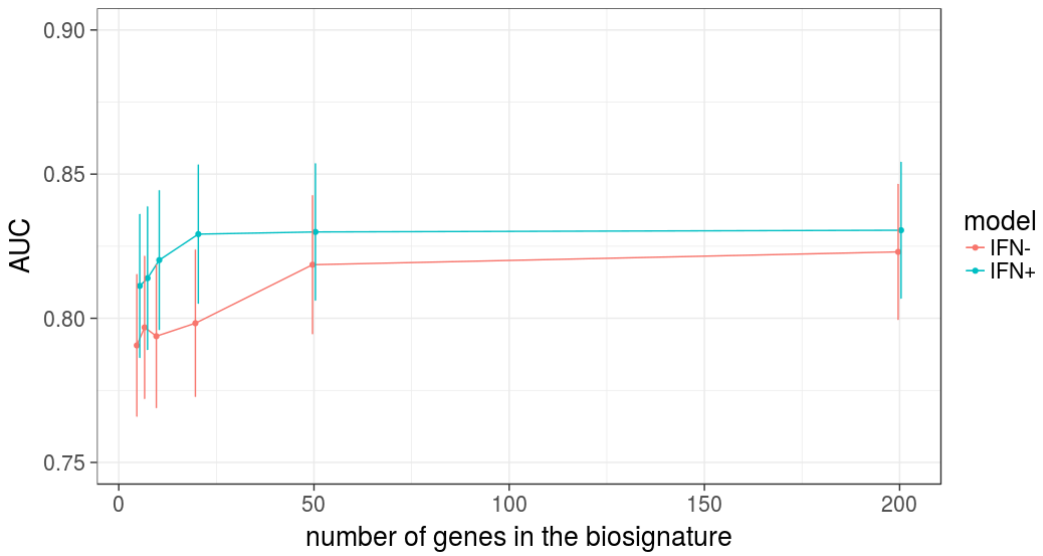


Figure 26 Dependence of the AUC of TB patients classification on the number of genes in the biosignature

The bars indicate 95% confidence intervals of calculated AUC.

I derived the IFN+ TB biosignature using the whole training MDS to build the RF model. Subsequently, 20 top genes sorted by variable importance were chosen and tested on the previously unused test MDS containing 20% of untouched samples from every study. Similarly, I derived the 50-gene signature of TB IFN- patients.

Table 9 Signature transcripts of IFN+ and IFN- TB patients

The transcript ENSEMBL IDs and corresponding HGNC gene names are given. The third column indicates the transcripts overlapping in the IFN+ and IFN- biosignature. The column "Kaforou 53 Transcript Signature overlap indicates the transcripts previously indicated as biosignature of TB vs non-TB by Kaforou et al. (Kaforou et al., 2013). The last column indicates transcripts which were indicated as IFN type I inducible or inducible by both IFN type I and type II by Interferome v2.0 database.

ENSEMBL ID	HGNC symbol	IFN-/IFN+ TB Signature overlap	Kaforou 53 Transcript Signature overlap	Present in IFN I or IFN I and II module
		IFN+ TB Biosignature		
ENSG00000002549	LAP3			+
ENSG00000070501	POLB	+	+	+
ENSG00000100911	PSME2			+
ENSG00000108387	SEPT4	+	+	+
ENSG00000108861	DUSP3		+	
ENSG00000120217	CD274			+
ENSG00000135148	TRAFD1			+
ENSG00000150337	FCGR1A	+	+	+
ENSG00000152223	EPG5			
ENSG00000152766	ANKRD22	+	+	+
ENSG00000154451	GBP5	+	+	+
ENSG00000162645	GBP2			+
ENSG00000163568	AIM2			+
ENSG00000168062	BATF2	+		+
ENSG00000168899	VAMP5	+		+
ENSG00000173369	C1QB			+
ENSG00000185338	SOCS1	+		+
ENSG00000225492	GBP1P1			+
ENSG00000225967	TAP2			
ENSG00000265531	FCGR1CP	+		
		IFN- Biosignature		
ENSG00000001084	GCLC			
ENSG00000003436	TFPI			
ENSG00000004939	SLC4A1			
ENSG00000070501	POLB	+	+	+
ENSG00000089057	SLC23A2			
ENSG00000090659	CD209			
ENSG00000100568	VTI1B			
ENSG00000108387	SEPT4	+	+	+
ENSG00000112640	PPP2R5D			
ENSG00000119906	SLF2			

ENSG00000126012	KDM5C			
ENSG00000128274	A4GALT			
ENSG00000129003	VPS13C			
ENSG00000137959	IFI44L			+
ENSG00000140105	WARS			+
ENSG00000140287	HDC			
ENSG00000145685	LHFPL2		+	
ENSG00000145936	KCNMB1			+
ENSG00000149131	SERPING1			+
ENSG00000150337	FCGR1A	+	+	+
ENSG00000152229	PSTPIP2			
ENSG00000152766	ANKRD22	+	+	+
ENSG00000154451	GBP5	+	+	+
ENSG00000159173	TNNI1			
ENSG00000161133	USP41			+
ENSG00000164330	EBF1		+	
ENSG00000165416	SUGT1			
ENSG00000167995	BEST1			
ENSG00000168062	BATF2	+		+
ENSG00000168899	VAMP5	+		+
ENSG00000174944	P2RY14			+
ENSG00000180185	FAHD1			
ENSG00000185338	SOCS1	+		+
ENSG00000186625	KATNA1			
ENSG00000187608	ISG15			+
ENSG00000188820	FAM26F		+	+
ENSG00000188938	FAM120AOS			
ENSG00000196141	SPATS2L			+
ENSG00000196961	AP2A1			
ENSG00000198019	FCGR1B	+	+	+
ENSG00000204257	HLA-DMA			
ENSG00000205730	ITPRIPL2			
ENSG00000211978	IGHV5-78			
ENSG00000226264	HLA-DMB			
ENSG00000228163	HLA-DPA1			
ENSG00000234154	HLA-DMB			
ENSG00000239329	HLA-DMB			
ENSG00000241394	HLA-DMA			
ENSG00000242574	HLA-DMB			
ENSG00000265531	FCGR1CP			

The classification of the samples in the test set based on the IFN+ and IFN- TB biosignatures was characterized by better overall sensitivity and specificity than the classification using the 10-fold cross validation (Figure 27). The IFN+ TB biosignature presented the best performance on classifying the IFN+ TB patients (AUC = 0.93, 95% CI = 0.89 – 0.96), however the overall performance of the TB

IFN+ biosignature (AUC 0.78, 95% CI = 0.73 – 0.84) as well as the stability of the model were worse than the performance of the IFN- biosignature (AUC = 0.84, 95% CI = 0.79 – 0.89). In both IFN+ and IFN- TB biosignatures there were genes overlapping with the 53 transcript signature of TB vs non-TB identified by Kaforou et al. (Kaforou et al., 2013). Interestingly, 18 out of 50 genes present in IFN- TB biosignature were the genes inducible by type I IFN.

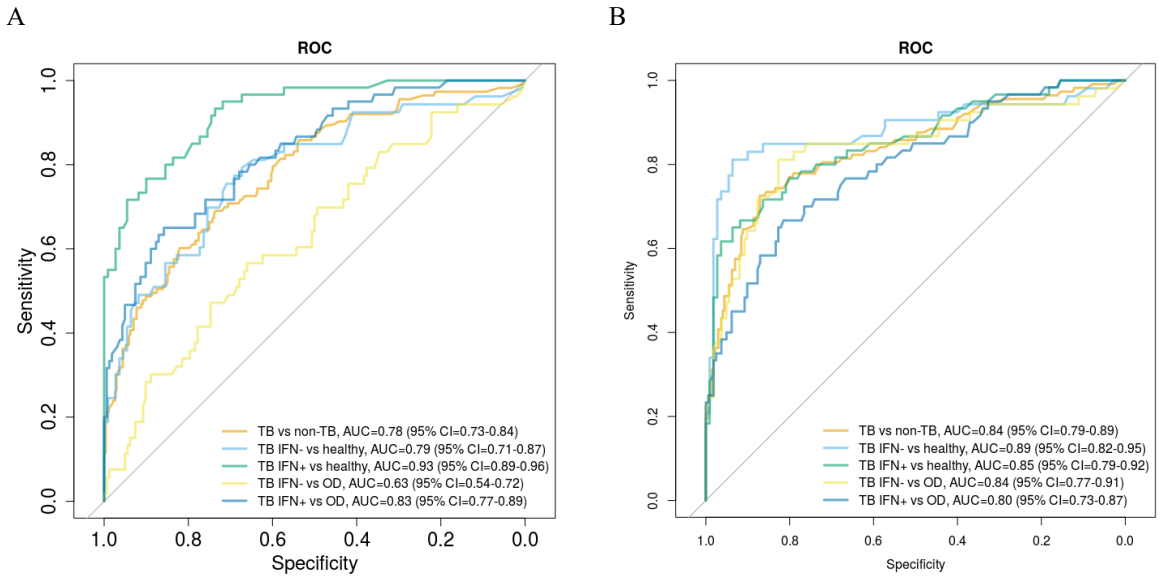


Figure 27 Performance of the TB biosignatures on the test MDS

The overall performance of the TB IFN- biosignature (B) is more robust than of the IFN+ TB biosignature (A; AUC = 0.84 for IFN- TB biosignature vs AUC = 0.78 for IFN+ TB biosignature). Additionally, the performance of the TB IFN- biosignature is more stable in identification of TB patients among various non-TB patient groups.

Altogether, the optimal biosignature sizes for IFN- and IFN+ TB were 50- or 20- transcripts, respectively. The performance of the biosignatures in identifying TB patients was even better on the test than on the training MDS. The IFN- based models were significantly more stable than the IFN+ based models and were characterized by high sensitivity and specificity for detection of all subgroups of TB patients in contrast to the IFN+ models which did not sensitively or specifically detect IFN- TB patients among OD patients.

3.14. PERFORMANCE OF THE TB IFN- AND TB IFN+ BIOSIGNATURES ON AN EXTERNAL DATASET FROM CHINA

To test the derived biosignatures I acquired an additional dataset from the WB of TB patients and HCs from China (Cai et al., 2014). After exclusion of the individuals who underwent anti-TB treatment the set contained 9 active TB cases, 6 LTBI and 6 uninfected control individuals. GSEA based identification of IFN+ and IFN- individuals showed that all the active TB cases from this dataset were IFN+. Nevertheless, I tested the performance of both the 20-transcript IFN+ TB biosignature and the 50-transcript IFN- TB biosignature for the classification of the TB cases in this dataset and visualized it using ROC curves. Since not all the transcripts from IFN- TB biosignature were measured in the dataset, I reduced the signature to 48 genes present in the dataset.

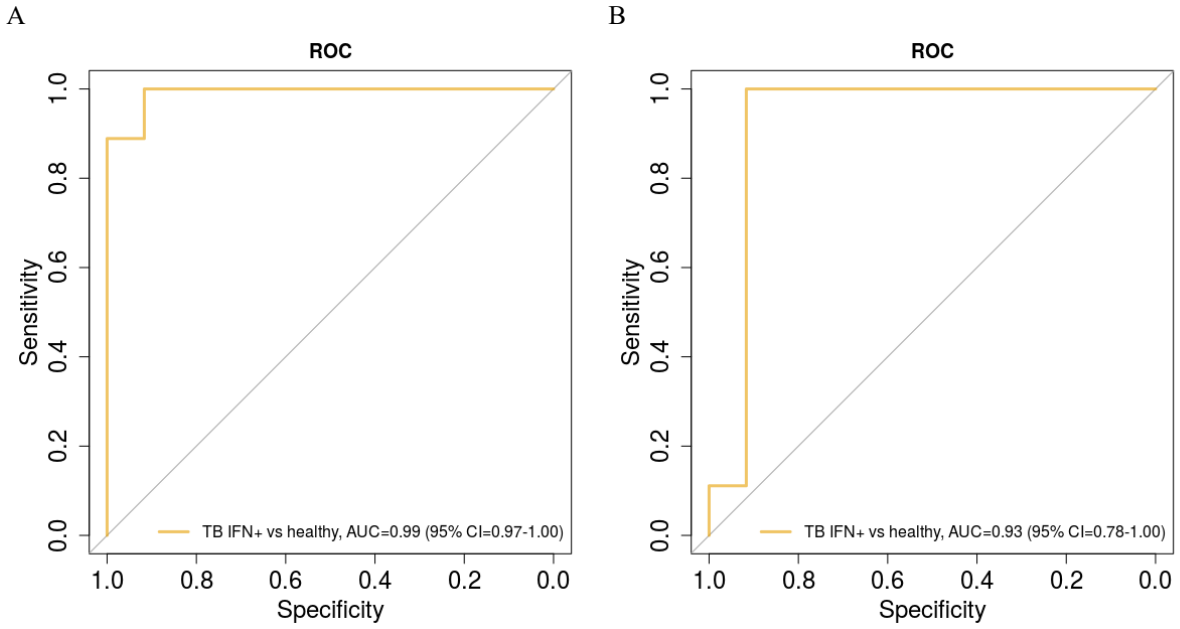


Figure 28 Performance of the TB biosignatures on the validation dataset from China

(A) 20-transcript IFN+ TB biosignature, and (B) 48-transcript IFN- TB biosignature were tested on the external validation dataset from China.

The TB patients from the external test dataset from China were classified with AUC of 0.99 (95% CI = 0.97 – 1.00) by the 20-transcript IFN+ TB biosignature and with AUC of 0.93 (95% CI = 0.78 – 1.00) by the 48-transcript IFN- TB biosignature. The biosignatures obtained based on the MDS were therefore sensitive and specific to detect TB also in an independent dataset (Figure 28).

3.15. PERFORMANCE OF THE TB IFN- AND TB IFN+ BIOSIGNATURES IN DIFFERENTIATING BETWEEN TB AND SARCOIDOSIS PATIENTS

In the previous tests and validation I have shown that the TB IFN+ and IFN- biosignatures identify TB patients among other patients and healthy subjects with high accuracy. Sarcoidosis is a disease that on the molecular level has been described to be hardly distinguishable from TB (Maertzdorf et al., 2012). I tested if the derived biosignatures are able to identify TB patients among patients suffering from sarcoidosis using dataset containing 45 TB and 49 sarcoidosis patients and 61 healthy individuals (Blankley, Graham, Turner, et al., 2016). GSEA based identification of IFN+ and IFN- individuals showed that all the active TB cases from this dataset were IFN+. Nevertheless, I tested the performance of both the 20-transcript IFN+ TB biosignature and the 50-transcript IFN- TB biosignature for the classification of the TB cases in this dataset and visualized it using ROC curves.

The TB patients from the external test dataset from London were classified with AUC of 0.9 (95% CI = 0.85 – 0.95) by the 20-transcript IFN+ TB biosignature and with AUC of 0.75 (95% CI = 0.67 – 0.82) by the 50-transcript IFN- TB biosignature (Figure 29). There was a significant difference in the performance of the two signatures, which was the largest in the sensitivity and specificity of detection of the TB patients among the sarcoidosis patients. Strikingly, the IFN- signature was not able to identify the IFN+ TB patients among the sepsis patients. This suggests that in the regulation of IFN signaling pathways is crucial for differentiation between TB and sarcoidosis, two diseases inducing similar gene expression patterns. This could also indicate that on the gene expression level the sarcoidosis patients are more similar to IFN- TB patients than IFN+ TB patients.

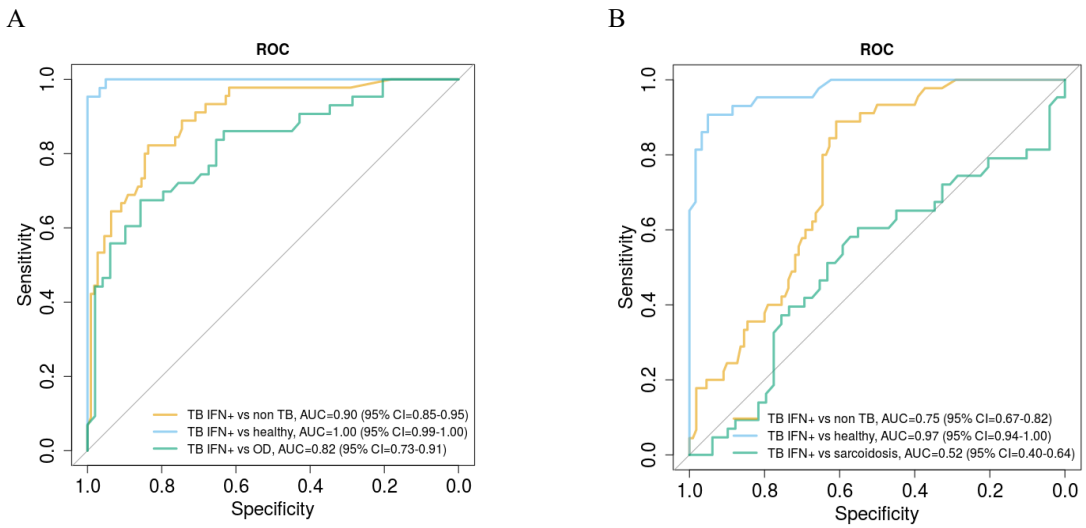


Figure 29 Performance of the TB biosignatures on the validation dataset including sarcoidosis patients

(A) 20-transcript IFN+ TB biosignature, and (B) 50-transcript IFN- TB biosignature were tested on the external validation dataset from London. The IFN- TB signature is not specific for detection of TB patients among the sarcoidosis patients.

3.16. VALIDATION OF THE METHODS ON SEPSIS DATASETS

I have acquired three datasets from sepsis patients and HCs in order to test the method identification of IFN- and IFN+ individuals and to check what performance will characterize biosignatures of IFN- and IFN+ sepsis patients derived with the presented above methods. Data acquisition, normalization and GSEA were performed as described in the Methods section. Only transcripts measured in TB MDS were included in the sepsis MDS in order to derive a biosignature that could be later tested on the TB MDS. The sepsis MDS was split into the training and test dataset. In the training Sepsis MDS there were 32 IFN- Sepsis patients, 46 IFN+ Sepsis patients and 44 HC samples. RF models were trained using 10-fold cross validation, separately on sepsis IFN+ individuals and HCs, and on sepsis IFN- individuals and HCs. To acquire signatures of comparable size corresponding with TB IFN+ and TB IFN- signatures, 20-transcript sepsis IFN+ and 50-transcript sepsis IFN- signatures were derived from the RF models and applied to the test set. The sensitivity and specificity of sepsis patients classification was visualized using ROC curves (Figure 30).

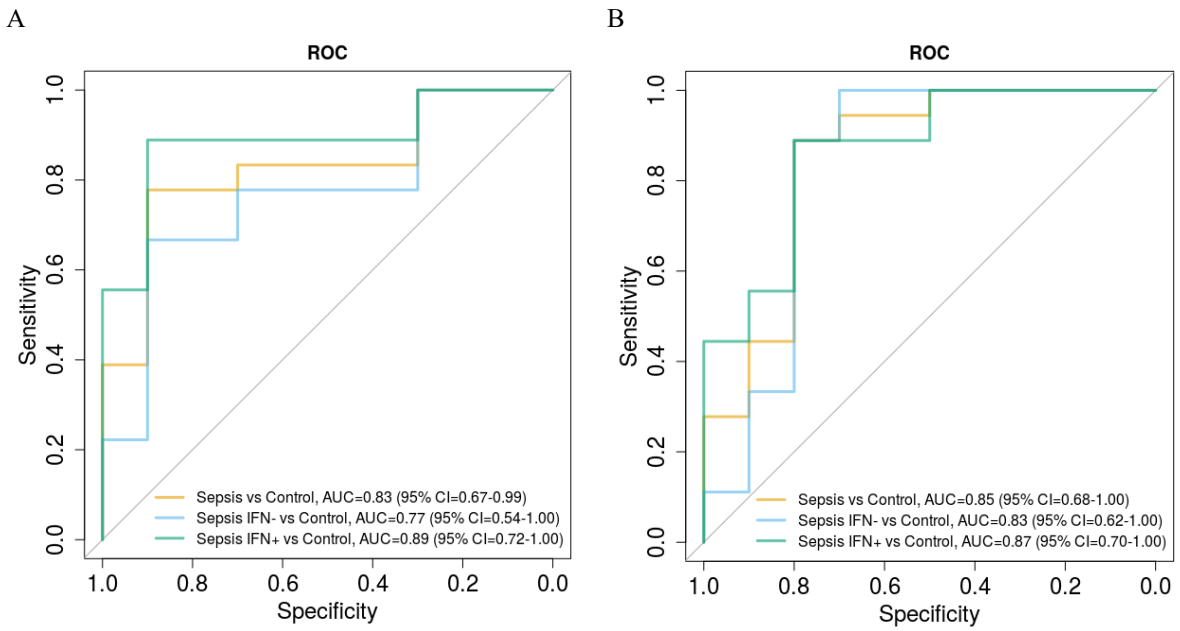


Figure 30 Performance of the sepsis biosignatures on the sepsis test MDS

(A) 20 transcript IFN+ sepsis biosignature and **(B)** 50 transcript IFN- sepsis biosignature were tested on the sepsis test set.

The 20-transcript sepsis IFN+ signature based models classified the sepsis patients in the test set with overall AUC of 0.83 (95% CI = 0.67 – 0.99) while the 50-transcript sepsis IFN- biosignature classified patients in the test set with even higher AUC of 0.85 (95% CI = 0.68 – 1.00).The 95% CI were broad due to small number of samples in the sepsis test MDS. The classification of IFN – sepsis patients was less sensitive and specific independent of the model used. There was only one transcript overlapping between the IFN+ and IFN- sepsis signatures and one between the sepsis IFN+ and TB IFN+ biosignature (Table 10). 4 out of 20 genes present in IFN+ sepsis biosignature and 7 out of 50

genes present in IFN- sepsis biosignature were classified as IFN type I inducible by Interferome v2.0 database. Some of the genes present in the sepsis IFN+ and sepsis IFN- biosignatures have been previously listed among significantly regulated genes in studies on sepsis or one of its causes - bacterial meningitis (Foell et al., 2013; Lill et al., 2013; Zhang et al., 2014).

Table 10 Biosignatures of the IFN+ and IFN- sepsis

ENSEMBL transcript identifiers and HGNC gene names are listed. There was no overlap between sepsis IFN+ and sepsis IFN- signature. Many of the genes present in IFN+ Sepsis signature have been previously reported in the studies on Sepsis or bacterial meningitis – examples of them are presented in the column "Previously listed" as the references to publications in which they have been listed. 6 out of 20 genes from IFN+ sepsis biosignature and 7 out of 50 genes from IFN- sepsis biosignature have been classified as IFN type I inducible by Interferome v2.0 database.

ENSEMBL ID	HGNC symbol	sepsis Signatures overlap	TB signatures overlap	Previously listed	Present in IFN I or IFN I and II module
IFN+ sepsis BIOSIGNATURE					
ENSG00000090376	IRAK3	-	-	(Lill et al., 2013)	+
ENSG00000104814	MAP4K1	-	-	(Priya et al., 2017)	
ENSG00000129682	FGF13	-	-	(Basu et al., 2011)	
ENSG00000131378	RFTN1	-	-		
ENSG00000135404	CD63	-	-	(Lill et al., 2013; Wan-Chung Hu, 2013)	
ENSG00000137767	SQOR	-	-		
ENSG00000138772	ANXA3	-	-	(Fiusa et al., 2014)	
ENSG00000150045	KLRF1	+	-	(Wan-Chung Hu, 2013)	
ENSG00000152766	ANKRD22	-	+		+
ENSG00000156414	TDRD9	-	-	(Davenport et al., 2016)	
ENSG00000159339	PADI4	-	-	(Lill et al., 2013)	
ENSG00000163754	GYG1	-	-	(Lill et al., 2013)	
ENSG00000166507	NDST2	-	-	(Oshima, Haeger, Hippensteel, Herson, & Schmidt, 2018)	
ENSG00000166527	CLEC4D	-	-		
ENSG00000183019	MCEMP1	-	-		
ENSG00000187554	TLR5	-	-	(Lill et al., 2013)	+
ENSG00000198814	GK	-	-		+
ENSG00000206379	FLOT1	-	-	(Lill et al., 2013)	
ENSG00000230143	FLOT1	-	-	(Lill et al., 2013)	
ENSG00000236271	FLOT1	-	-	(Lill et al., 2013)	
IFN- sepsis BIOSIGNATURE					
ENSG00000008130	NADK	-	-		+
ENSG00000010244	ZNF207	-	-		
ENSG00000023516	AKAP11	-	-		
ENSG00000065154	OAT	-	-		+
ENSG00000085788	DDHD2	-	-		
ENSG00000100207	TCF20	-	-		
ENSG00000101000	PROCR	-	-	(Schouten et al., 2014)	
ENSG00000101665	SMAD7	-	-		

ENSG00000103174	NAGPA	-	-		
ENSG00000105607	GCDH	-	-		
ENSG00000108106	UBE2S	-	-		+
ENSG00000109063	MYH3	-	-		
ENSG00000110108	TMEM109	-	-		
ENSG00000110880	CORO1C	-	-		
ENSG00000111671	SPSB2	-	-		
ENSG00000113368	LMNB1	-	-	(Zhang et al., 2014)	+
ENSG00000114251	WNT5A	-	-		
ENSG00000117133	RPF1	-	-		
ENSG00000118418	HMGN3	-	-		+
ENSG00000123159	GIPC1	-	-		
ENSG00000129187	DCTD	-	-		
ENSG00000129351	ILF3	-	-		
ENSG00000135241	PNPLA8	-	-		
ENSG00000143889	HNRNPLL	-	-		
ENSG00000150045	KLRF1	+	-	(Wan-Chung Hu, 2013)	
ENSG00000155256	ZFYVE27	-	-		
ENSG00000161944	ASGR2	-	-		
ENSG00000163251	FZD5	-	-		+
ENSG00000164047	CAMP	-	-		
ENSG00000173372	C1QA	-	-	(Wan-Chung Hu, 2013)	+
ENSG00000174695	TMEM167A	-	-		
ENSG00000177479	ARIH2	-	-		
ENSG00000184922	FMNL1	-	-		
ENSG00000185905	C16orf54	-	-		
ENSG00000187475	HIST1H1T	-	-		
ENSG00000187764	SEMA4D	-	-		
ENSG00000188636	RTL6	-	-		
ENSG00000196653	ZNF502	-	-		
ENSG00000198018	ENTPD7	-	-		
ENSG00000198258	UBL5	-	-		
ENSG00000198736	MSRB1	-	-		
ENSG00000203666	EFCAB2	-	-		
ENSG00000239961	LILRA4	-	-		
ENSG00000241468	ATP5MF	-	-		
ENSG00000275596	KIR2DL5A	-	-		
ENSG00000276068	NAIP	-	-		
ENSG00000276461	TCF20	-	-		
ENSG00000277667	TMC4	-	-		
ENSG00000278481	KIR2DL5B	-	-		
ENSG00000281794	MUC20-OT1	-	-		

In summary, the GSEA-based method of identification of IFN+ and IFN- subgroups of patients detected those two subgroups among sepsis patients. The derived sepsis biosignatures presented

variable performance: with high sensitivity and specificity of sepsis patient classification based on IFN+ patients and lower performance based on IFN- sepsis patients.

3.17. TESTING TB BIOSIGNATURES ON SEPSIS PATIENTS

Various infectious diseases share many common mechanisms of the response to the invading pathogen. For example, even though TB and sepsis involve different disease mechanisms, they both are characterized by strong inflammatory and IFN response. To investigate, how specific to TB are the obtained TB IFN- and TB IFN+ transcript signatures, I tested their performance on identification of the sepsis patients using sepsis test MDS. I used the TB IFN+ and TB IFN- RF models to predict which donors in sepsis test MDS are sick.

The AUC of the classification of sepsis patients from healthy based on 20-transcript TB IFN+ signature equaled 0.62 (95% CI = 0.40-0.83) which was significantly less than in the case of the 20-transcript sepsis IFN+ signature and not significantly better than random prediction (Figure 31). The AUC of the classification of sepsis patients from healthy based on 50-transcript TB IFN- signature was even worse and equaled 0.54 (95% CI = 0.31-0.76). In the case of both signatures the sensitivity and specificity of detecting IFN- sepsis patients was lower than that of detecting IFN + sepsis patients which suggests that the common IFN response in IFN+ TB and IFN+ sepsis patients share part of their gene regulation profile, which is not disease but IFN specific.

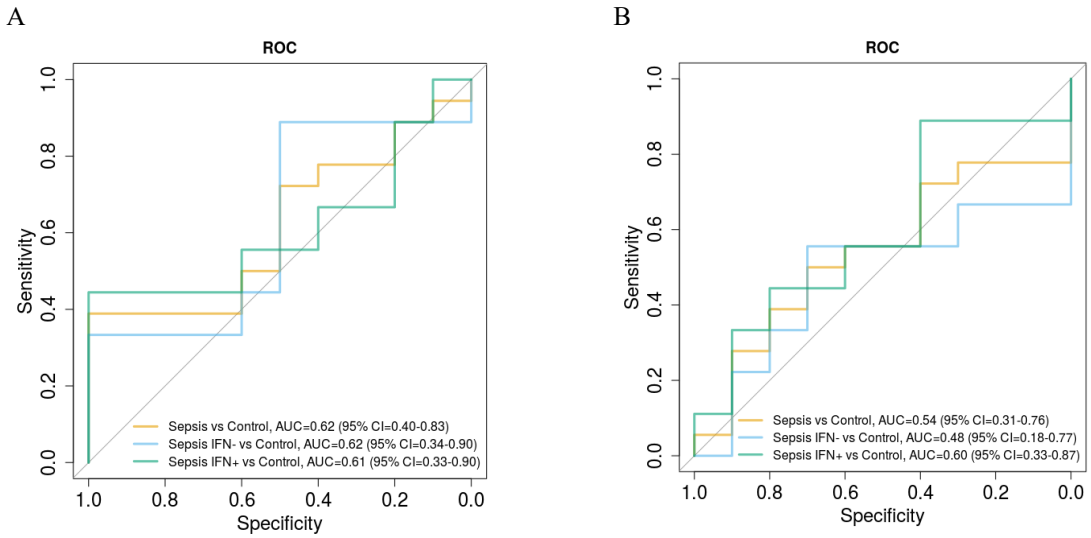


Figure 31 Performance of the TB biosignatures on the sepsis test MDS

(A) 20 transcript IFN+ TB biosignature and (B) 50 transcript IFN- TB biosignature were tested on the sepsis test set. The TB IFN+ and IFN- signatures are not sensitive and specific towards detection of sepsis.

I observed that the IFN- TB signature presents higher specificity for the detection of TB which is likely due to the fact that it is not convoluted with the IFN pathway activation occurring in many infectious diseases.

3.18. TESTING SEPSIS BIOSIGNATURES ON TB PATIENTS

I subsequently tested the performance of the obtained sepsis biosignatures on the TB test MDS. The AUC of the classification of TB patients from non-TB based on the 20-transcript TB IFN+ signature equaled 0.69 (95% CI = 0.63-0.75) which was significantly less than in the case of the 20-transcript TB IFN+ signature (Figure 32). The AUC of the classification of TB patients from non-TB based on the 50-transcript TB IFN- signature was similar and equaled 0.65 (95% CI = 0.59-0.71). In the case of both signatures the sensitivity and specificity of detecting IFN- TB patients from OD patients was not significantly different than 0.5. The sensitivity and specificity of detecting IFN+ TB patients was in both cases highest, which indicates that the IFN signaling convoluted the transcriptional differences between the two diseases.

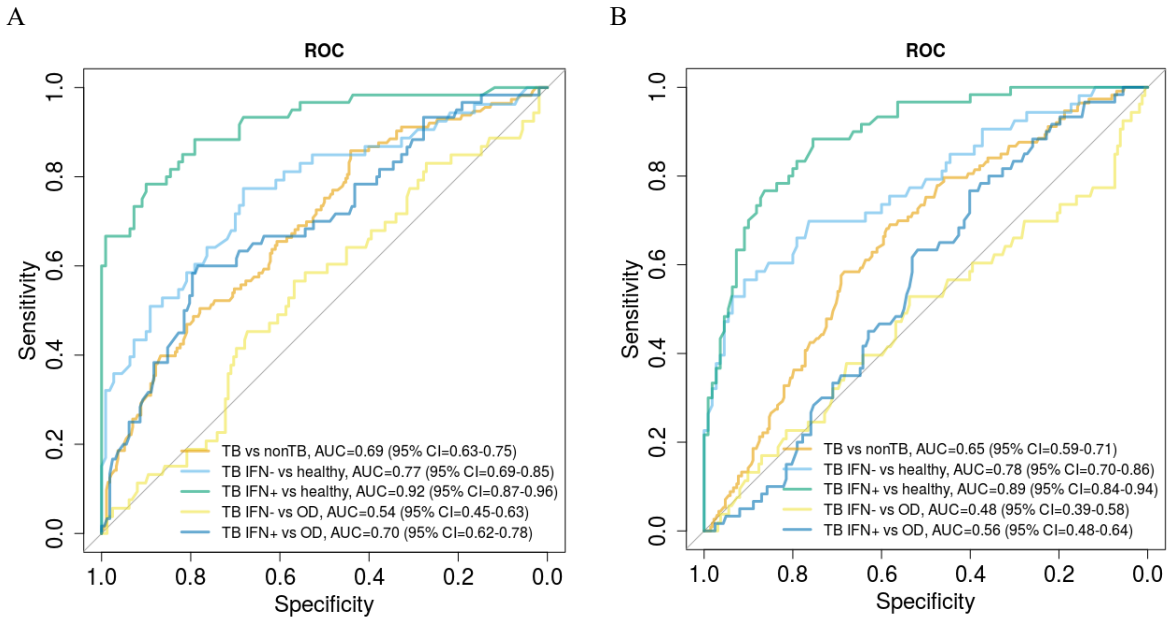


Figure 32 Performance of the sepsis biosignatures on the TB test MDS

(A) 20 transcript IFN+ sepsis biosignature and **(B)** 50 transcript IFN- sepsis biosignature were tested on the TB test set. The overall performance of the signatures is better than random.

3.19. PROFILES OF IMMUNE RESPONSE IN TB PATIENTS

Not only IFN response but also various other modules presented variable enrichment among TB patients compared to healthy. To investigate how the gene expression profiles in different modules differ between patients I performed correlation analysis of the gene expression in transcriptomic modules. Since every module consists of multiple genes, I calculated eigengenes representing the gene expression in modules and visualized the correlation between them on the heatmap (Figure 33).

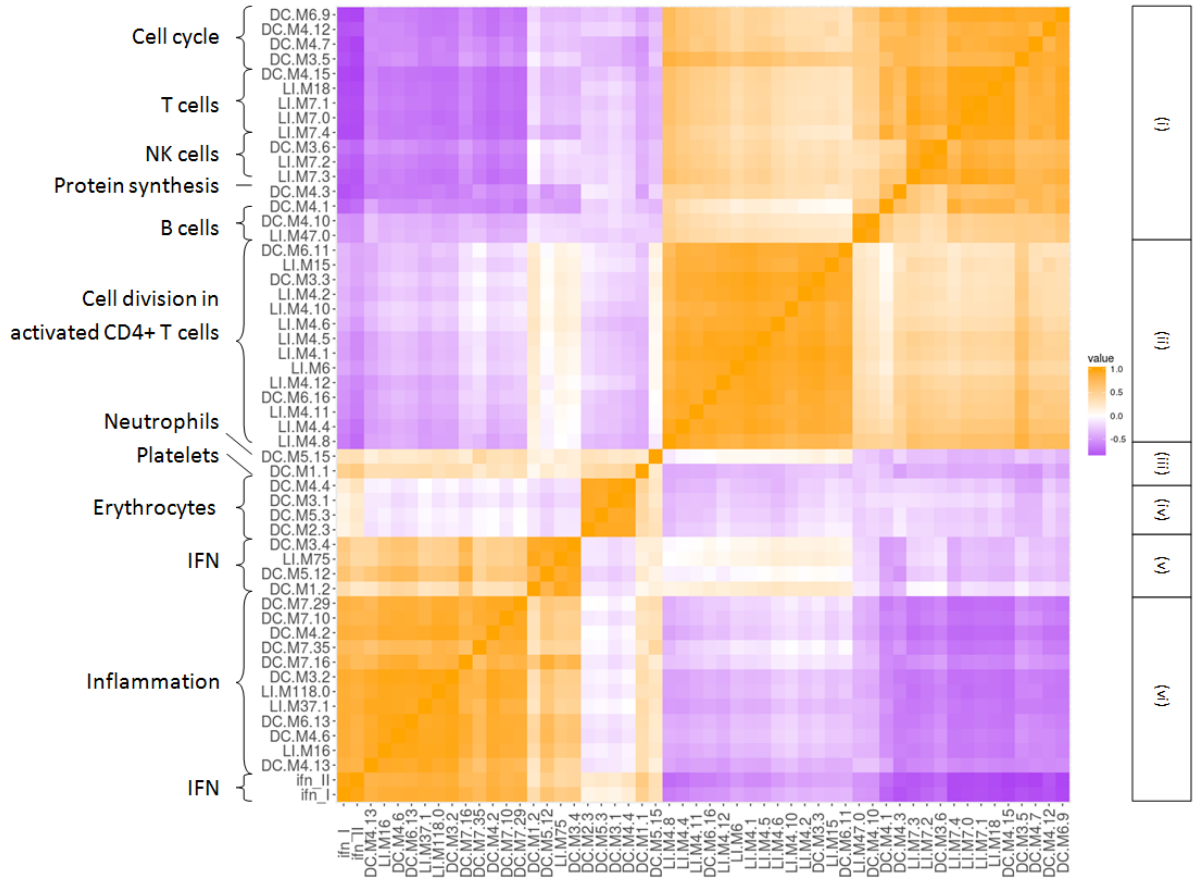


Figure 33 Heatmap of correlations of gene expression in modules

Every row and column represent the module which ID is described in the row name. The modules are clustered based on the correlation and groups of modules are annotated on the left according to the dominating term in the module cluster. The six described major clusters are marked on the right.

The correlation matrix indicated 6 different patterns of gene expression regulation:

(i) Strong up-regulation of genes related to cell cycle, T-cells, NK-cells and B-cells correlated with strong down-regulation of genes related to IFN type I and type II response and inflammatory processes, and mild down-regulation of genes related to erythrocytes, platelets, and neutrophils, as well as mild up-regulation of genes related to cell division processes in activated CD4+ T-cells;

(ii) Strong up-regulation of genes related to cell division processes in activated CD4+ T cells correlated with mild down-regulation in the modules related to inflammatory processes and erythrocytes and mild up-regulation of genes related to NK-cells, T-cells, IFN type I and II response and cell cycle;

(iii) Strong up-regulation of genes related to erythrocytes corresponded with mild up-regulation of genes related to neutrophils and platelets, as well as mild down-regulation of genes related to cell division in activated CD4+ T cells, inflammation, cell cycle, T-cells, B-cells and NK-cells;

(iv) Strong up-regulation of genes related to neutrophils and platelets, up-regulation of genes related to type I and type II IFN signaling, inflammation, erythrocytes and down-regulation of genes related to cell division in activated CD4+ T-cells, protein synthesis, NK-cells, T-cells and cell cycle;

(v) Strong up-regulation of IFN type I and type II IFN signaling genes correlated with up-regulation of genes related to inflammatory processes, strong down-regulation of genes related to protein synthesis and T-cells and mild down-regulation of genes related to cell cycle and NK-cells;

(vi) Strong up-regulation of genes related to inflammatory response correlated with up-regulation of genes related to IFN type I and II signaling, neutrophils, platelets and down-regulation of genes related to NK cells, T-cells, B-cells, cell cycle and mild down-regulation of genes related to cell division processes in CD4+ T cells.

In summary, I identified six different patterns of immune response against TB presented by the subsets of TB patients.

3.20. DISEASE RISK SCORE DOES NOT CORRESPOND TO INTERFERON STATUS

To assess if the IFN positivity status correlates in TB patients with disease risk score (DRS) calculated as proposed by Kaforou et al. (2013), I classified each study participant as "DRS TB" or "DRS healthy" based on the calculated DRS score. The DRS assigned 84% of IFN- patients as "not TB" and 51% of IFN+ as "not TB". The pairwise comparison of proportions using a Fisher's exact test with Bonferroni correction indicated significant difference between the proportions of patients classified as "TB" and "not TB" in the IFN+ and IFN- group (Figure 34).

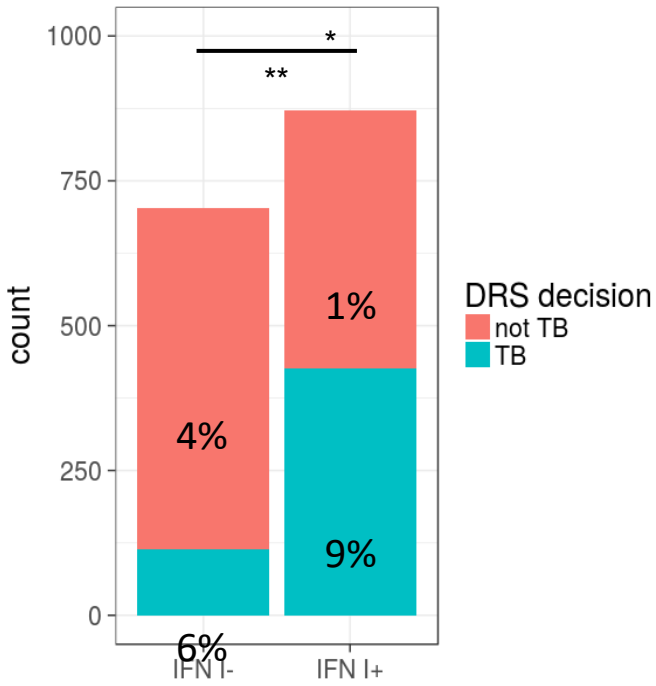


Figure 34 Proportions of the IFN+ and IFN- individuals from MDS assigned as "TB" and "not TB" by the DRS

The TB patients are marked green. The non-TB individuals are marked red. The pairwise comparison of the proportions using the exact Fisher's test with Bonferroni correction was significant.

Closer inspection of the classification of the patients by DRS score indicated that depending on the real status of the individual (TB, OD or healthy), the DRS was characterized by varying performance in correctly classifying the patients as "TB" and "not TB". In every group of individuals (TB, healthy, OD) samples classified as "not TB" as well as "TB" consisted of a mixed population of IFN- and IFN+ individuals. The DRS classified correctly 95% of healthy (including LTB) patients as "not TB". Among the patients suffering from other pulmonary diseases it wrongly assigned 33% of them as "TB" and among the TB patients it assigned 27% as "not TB". Half of the TB patients assigned as "not TB" were IFN-. Comparison of the classification as "not TB" or "TB" by the DRS was significantly related to the assignment of "IFN-" and "IFN+" status in all the groups. This suggests that the DRS is dependent on the strength of IFN response in the patients (Figure 35).

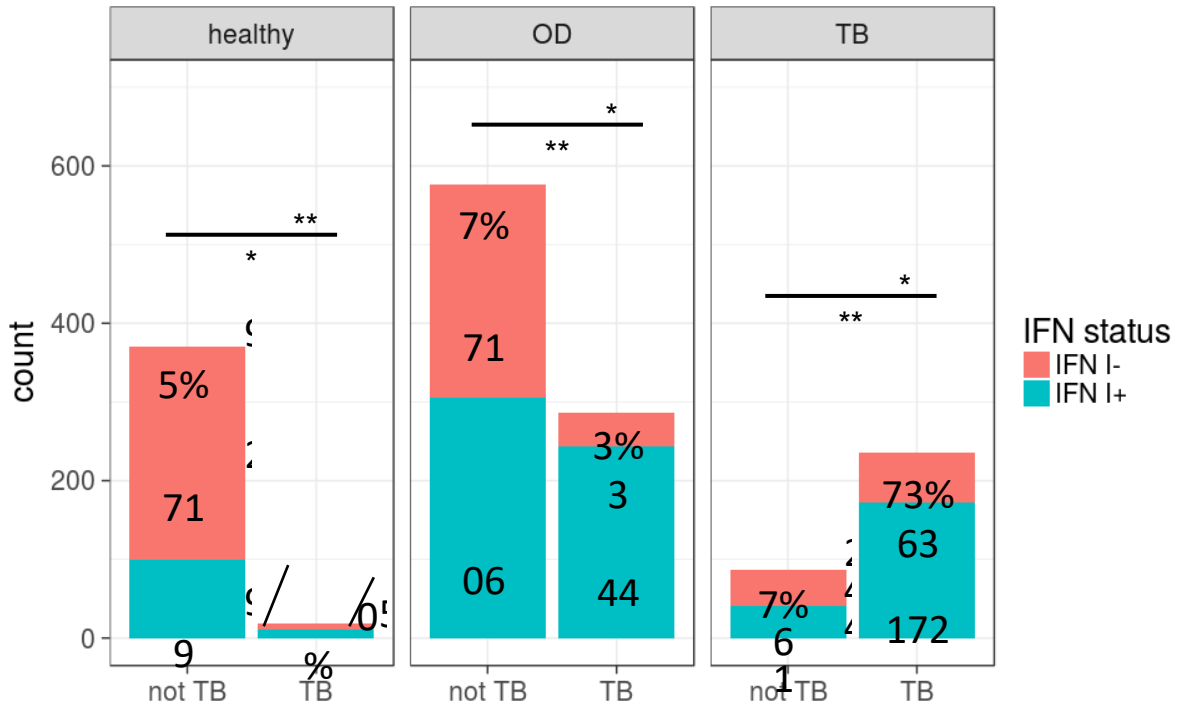


Figure 35 Fraction of IFN- and IFN+ samples among individuals classified as non-TB and TB by DRS in the three groups of donors: healthy, OD and TB

The IFN- patients are marked green. The IFN+ individuals are marked red. The pairwise comparison of the proportions using the exact Fisher's test with Bonferroni correction was significant

10 out of 53 transcripts in the 53-gene "TB vs non TB" signature of Kaforou et al. (Kaforou et al., 2013) were assigned as IFN type I inducible by the Interferome v2.0 database. The bias towards assigning of the IFN+ patients as "TB" could be explained by two hypotheses: either the DRS is strongly influenced by the presence of the 10 IFN type I genes in the 53 transcript used to create the DRS, or the IFN- TB patients present too small extent of gene regulation upon TB to be correctly detected by the DRS. In both cases, considering the IFN status of the patients proves important for the diagnosis based on the DRS.

3.21. INFLUENCE OF TIME POST INFECTION ON INTERFERON STATUS

Generally, it is not known how much time has passed since the primary infection until the moment of diagnosis of TB in patients. Therefore, we cannot exclude that the differences seen in IFN status of TB patients are results of the different disease development stage with IFN responses becoming stronger with the ongoing infection. To investigate this hypothesis I acquired a dataset from blood of 38 macaques (*Macaca fascicularis*, also referred to as cynomolgus macaques) infected with Mtb and followed in 11 serial time points for 6 months after the infection (Gideon et al., 2016). At each time point the blood transcriptome of the macaques was quantified with microarrays and the infection

outcome was defined using clinical definitions of active TB and LTBI, as well as on the basis of total lung inflammation measured as levels of [^{18}F] fluorodeoxyglucose (FDG), a surrogate marker for disease severity in lungs, measured by positron emission tomography - computed tomography (PET-CT). Out of the 38 animals, 16 developed active TB and 22 remained latently infected. I calculated z-scores for the gene expression as described before and GSE with the type I IFN module set for individual macaques at every time point. I observed that the peak of type I IFN response fell between the 20th and 42nd day after infection (Figure 36). Interestingly, there was no clear pattern in IFN type I module enrichment distinguishing the macaques which developed active TB from those remaining latently infected. There was a trend towards long lasting IFN type I response in the animals presenting active TB: 75% of them still presented with enrichment (p-value lower than $10^{-4.3}$) at day 56 p.i., whereas among the latently infected animals it was presented in only 27% of the animals. On the other hand, among the latently infected animals the enrichment in type I IFN response was observed earlier, e.g. at the day 3 p.i..

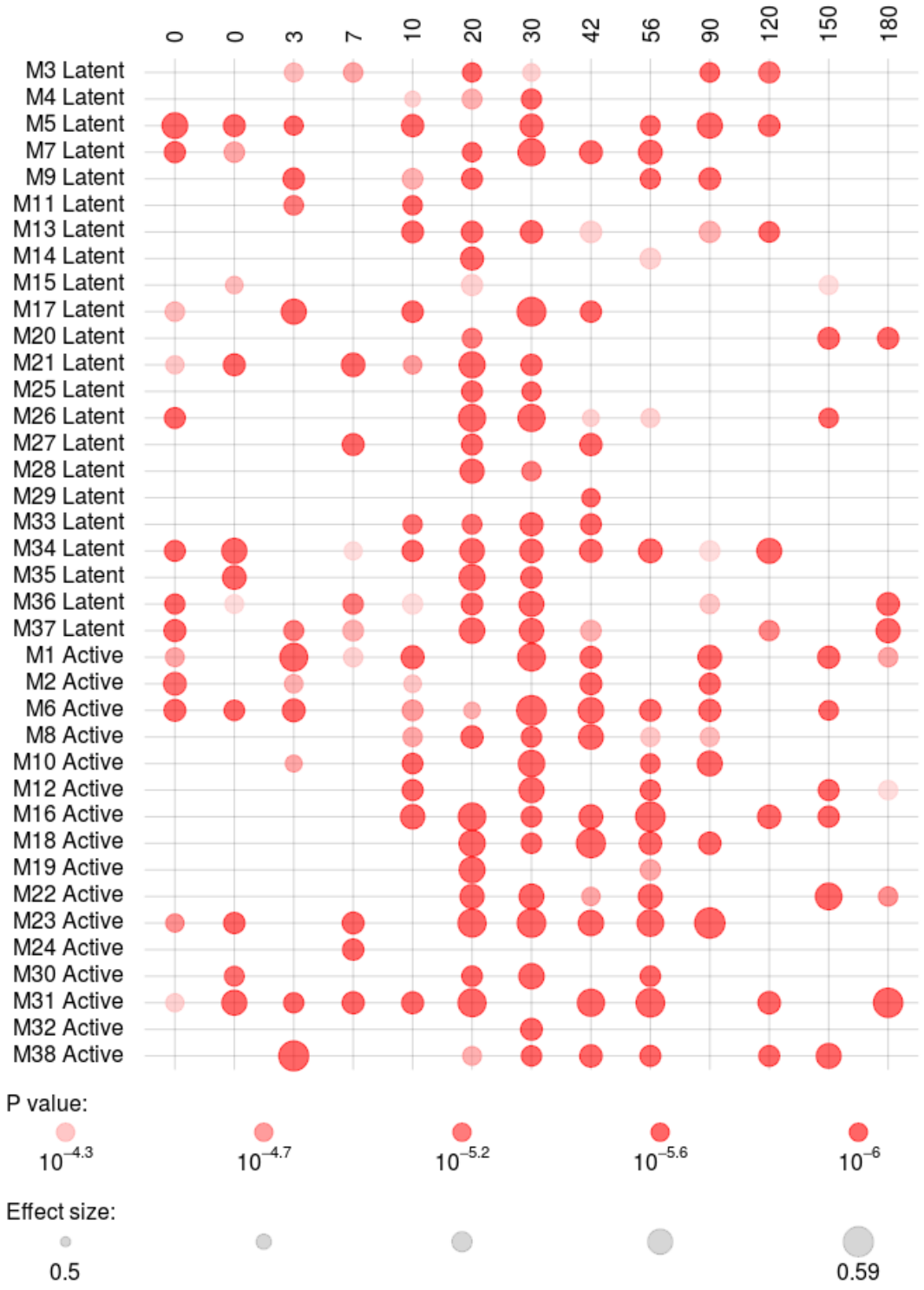


Figure 36 Enrichment of the "IFN type I" module in the individual macaques over the time pre- and post infection

The enrichment in these modules is significant in both LTB and TB macaques.

In conclusion, the study with controlled time p.i. suggested that all Mtb infected animals develop IFN type I response but the strength and time of the response is heterogeneous between individuals.

4. CHAPTER 4: IDENTIFICATION OF CONCORDANT AND DISCORDANT IMMUNE RESPONSES TO TUBERCULOSIS IN MOUSE AND MAN

Here, I introduce a method which allows identifying highly concordantly as well as highly discordantly regulated gene sets between two organisms. The method is based on measuring concordance using directionality of change weighted by the magnitude of gene expression change in two heterologous datasets (for example, human and murine) and associated precision of its estimate. To this end, the approach combines a novel measure of similarity with GSEA. To validate this approach, I identified modules of genes concordantly and discordantly expressed in WB during TB in human populations from different regions and two different murine TB models. I then verified whether the differences found in WB are present also in human and murine macrophages.

4.1. ABSTRACT

Understanding of human immune response to infection is in a large extent based on knowledge derived from the mouse model system. However, there are major differences between the disease outcome and symptoms in man and experimental mouse strains, conditioned by large evolutionary distance, individual variability which in mouse is influenced by inbreeding and the fact that mouse is not a natural host of many human pathogens.

I propose a novel data integration approach which identifies concordant and discordant gene expression patterns of the immune responses in heterologous datasets: disco.score. The method accounts for the directionality and magnitude of the expression change of every gene in stimulated *vs* unstimulated groups of samples. The main assumption is that parts of the immune response remain conserved between species, while other aspects have diverged over time of evolution.

Using the publicly available datasets as well as datasets collected by my colleagues form MPIIB Department of Immunology, Lisa Scheuermann and Anca Dorhoi, I compared human and murine transcriptional responses to Mtb infection in WB. In a complementary approach I also compared responses of macrophages from mouse and man to the Mtb. The results indicate profound differences between regulation of innate and adaptive immunity in man and mouse upon Mtb infection. I characterized differential regulation of T-cell related genes corresponding to the differences in phenotype between TB high and low susceptible mouse strains and identified the time point of 21 days p.i. of mice as best reflection of transcriptional responses in the studied human cohorts.

The implemented approach facilitates the choice of an appropriate animal model for studies of the human immune response to a particular disease and provides the basis for better understanding of differences leading to success or failure in translation of laboratory findings to clinical trials.

The study was published in September 2017 in Scientific Reports (Domaszewska et al., 2017).

4.2. COMPARABLE DATASET ACQUISITION

I calculated differential gene expression between WB transcriptome profiles of (i) TB patients and HCs; (ii) Mtb infected and uninfected 129S2 and C57BL/6 mice as well as between transcriptional profiles of Mtb infected and uninfected human and murine macrophages.

I created a list of possible comparisons between the human and mouse samples, separately for WB and macrophage samples, and identified orthologous gene pairs between human and murine genes for each such comparison (Table 11). Only the genes having 1:1 orthologs assigned by species interlinking in the Ensembl database, where homology predictions are generated by implementing maximum likelihood phylogenetic gene trees (Vilella et al., 2009) were included. Exclusion of potential in-paralogs and genes without mapped orthologs resulted in truncated list of genes remaining in every comparison.

Table 11 Characteristics of the performed mouse-human comparisons

The table contains IDs of the comparison as described in the Methods section (Table 6), number of genes represented on the human microarray from each comparison, number of genes represented on the murine microarray from each comparison and the identified 1:1 orthologous gene number.

Comparison IDs	# genes represented on human microarray platform	#genes represented on murine microarray platform	# 1:1 orthologs
1, 2, 7, 8, 13, 14, 19, 20	19743	21662	14712
3- 6, 9- 12, 15- 18, 21- 24	21714	21662	15004
25	19743	19946	14314
26	19743	20665	13881
27	20477	19946	11695
28	20477	20665	11630
29	20477	19946	11688
30	20477	20665	11409
31	20477	19946	11694
32	20477	20665	11416
33	19743	20665	13885
34	20477	20665	11417
35	20477	20665	11412
36	20477	20665	11419

4.3. CORRELATION OF THE ACQUIRED DATASETS

I tested the methods described in the studies of Seok et al. and Takao and Miyakawa (Seok et al., 2013; Takao & Miyakawa, 2014) to investigate how the human gene expression regulation upon TB is mimicked by murine gene expression regulation using the two presented correlation approaches. According to the method presented by Seok et al., I calculated the squared Pearson's correlation coefficients (r^2) of the fold changes of all significantly differentially expressed gene pairs (p-value <0.05). No significant correlation in the gene expression of human and murine WB or macrophage transcriptomic profiles upon Mtb infection was detected as indicated by the obtained r^2 values which were lower than 0.1 (Table 12). The Spearman's rank correlation coefficients (r) calculated according to the method described by Takao and Miyakawa resulted in maximal value of 0.559 for the comparison #33; however, the criterion of including only the genes significantly regulated in both species resulted in a minute number of genes included in the comparison, i.e. 101 out of 13,885 orthologous gene pairs in this specific comparison. The previously described correlation-based approaches tested in performed comparisons did not answer the question of the level of similarity of gene expression regulation in the investigated mouse models and human TB.

Table 12. Results of the correlation-based comparisons of the murine and human datasets

The comparison ID, number of identified 1:1 orthologs between murine and human genes, calculated squared Pearson's correlation coefficient (r^2) and number of genes included in the calculation according to the criteria described by (Seok et al., 2013), calculated Spearman's correlation coefficient (r) and number of genes included in the calculation according to the criteria described by (Takao & Miyakawa, 2014) are given.

Comparison ID	# 1:1 orthologs	r^2	included genes (Seok)	r	included genes (Takao)
1	14712	0.001	5709	0.05	221
2	14712	0.003	6368	0.163	699
3	15004	0.003	4381	0.238	211
4	15004	0.023	5212	0.343	704
5	15004	0	5570	0.081	251
6	15004	0.002	6332	0.067	813
7	14712	0.024	5658	-0.215	275
8	14712	0	5798	-0.064	289
9	15004	0.003	4399	-0.042	228
10	15004	0.004	4483	0.037	254
11	15004	0.007	5542	-0.156	314
12	15004	0	5650	-0.089	316

13	14712	0.001	6720	0.076	1065
14	14712	0	7019	0.148	1248
15	15004	0.002	5764	0.152	1008
16	15004	0.012	5877	0.268	1066
17	15004	0	6754	0.081	1247
18	15004	0.006	6905	0.237	1267
19	14712	0	10334	0.08	3772
20	14712	0.009	10409	0.231	3519
21	15004	0.009	1000	0.218	7979
22	15004	0.033	9708	0.363	1698
23	15004	0.004	10535	0.165	3753
24	15004	0.025	10235	0.331	3400
25	14314	0.031	10395	0.296	3821
26	13881	0.042	5779	0.417	488
27	11695	0.032	4945	0.314	159
28	11630	0.068	4352	0.242	135
29	11688	0.052	4972	0.369	214
30	11409	0.052	4582	0.374	162
31	11694	0.061	4949	0.468	158
32	11416	0.081	4542	0.367	126
33	13885	0	5739	0.559	101
34	11417	0.002	375	0.119	24
35	11412	0.012	451	0.319	21
36	11419	0.021	373	0.177	22

4.4. GENE SET ENRICHMENT ANALYSIS

In the next step I calculated GSE based on gene modules from Molecular Signatures Database (MSigDB; Subramanian et al., 2005), gene modules created by Chaussabel et al., Godec et al. and Li et al. (Chaussabel et al., 2008; Godec et al., 2016; Li et al., 2014). The module sets from Li et al. (2014) and Chaussabel et al. (2008) were particularly informative because of their thorough biological annotation and focus on immune response elements. Therefore I used them in further described analysis referring to them as “immune modules”. GSE testing on the genes sorted by increasing p-value for differential regulation in each comparison resulted in lists of significantly enriched transcriptional

modules which varied in length, being most abundant for the human WB samples and mouse WB samples from the time points of 14 and 21 days p.i (Figure 37). The innate immunity modules typically enriched among TB patients (as shown in Chapter 3), like “antiviral IFN signature”, “enriched in neutrophils” or “enriched in monocytes”, were present among the significantly enriched modules in all the samples and multiple other modules enriched in both datasets were overlapping, indicating that there might be functionally related sets of genes regulated concordantly in spite of the lack of significant correlation between the datasets.

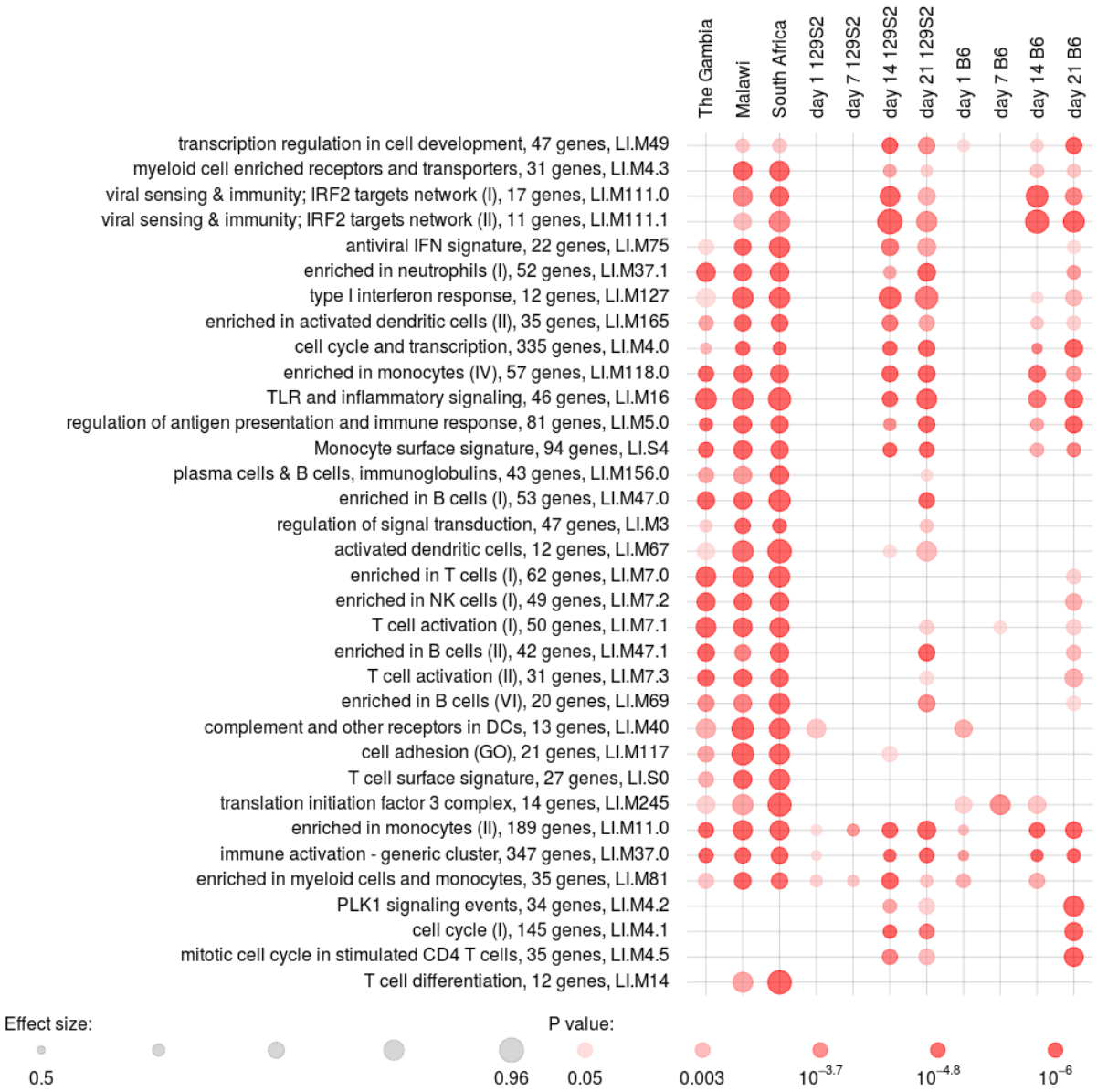


Figure 37 Gene expression patterns in the investigated human cohorts and murine WB from the 129S2 and C57BL/6 mice in days 1, 7, 14 and 21 p.i.

P-value of module enrichment is illustrated by the intensity of the color and the effect size by the size of the dot. There were no overlapping concordant modules present in these comparisons. The modules are described by the titles followed by the original number of genes in module and ID. Module IDs correspond to modules IDs in R package tmod (Domaszewska et al., 2017).

4.5. INTRODUCTION OF DISCO.SCORE

Since data heterogeneity did not allow meaningful statistical analysis of this phenomenon with currently available methods, I developed a measure to assess the similarity of gene expression regulation for each orthologous gene pair in two datasets. The following criteria were used to quantify this effect:

- magnitude of gene expression change (effect size)
- significance of gene expression change
- direction of gene expression change

I combined them into a single mathematical equation expressing a score of concordance/discordance in gene expression regulation between two datasets. The score, which I termed `disco.score` increases proportionally to both human and murine log_2FC increase (or decreases analogously), increases with the decrease of summed p-values of genes in pair, and has negative sign if the expression change has opposite direction (Equation 5, page 62).

The theoretical distribution of disco.score depending on log_2FC values in both datasets is illustrated in the (Figure 38).

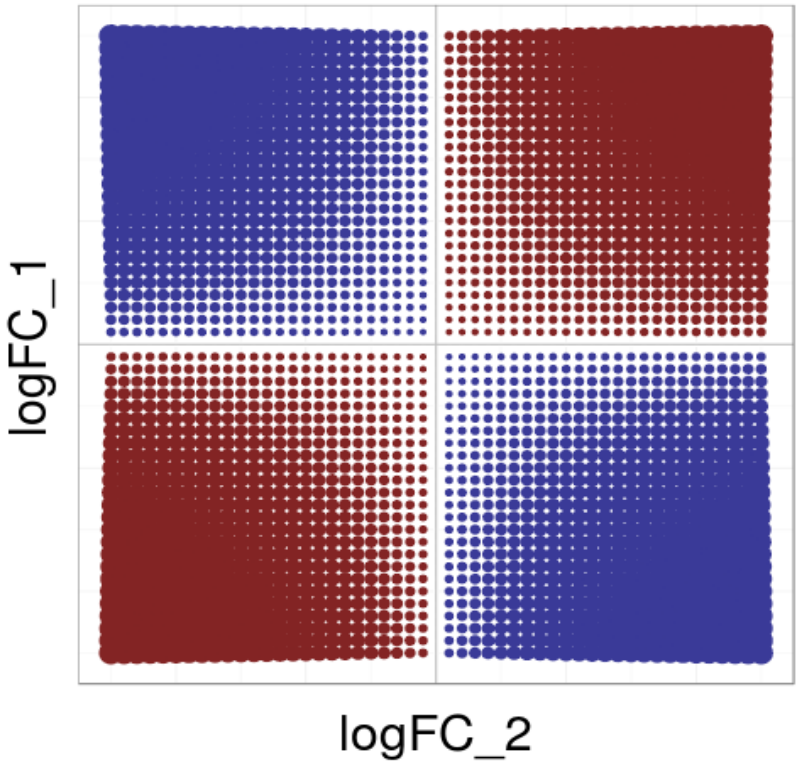


Figure 38 Theoretical distribution of disco.score function depending on log_2FC values of both species

Increasing intensity of the red color indicates increase in disco.score and illustrates higher degree of similarity between human and murine gene expression. Increasing intensity of the blue color indicates decrease in negative disco.score and a higher degree of dissimilarity between human and murine gene expression.

Using such defined formula I calculated disco.score for each pair of orthologous genes in every comparison. Then, I performed GSE on the list of genes from the two datasets ranked by decreasing disco.score to identify similar elements of immune response in every dataset pair. Moreover, using the list of genes sorted by increasing disco.score I identified the most dissimilar elements of immune response between each dataset pair. I termed the modules enriched in the dataset sorted by decreasing disco.score `concordant` and those enriched in dataset sorted by increasing disco.score `discordant` (Figure 39).

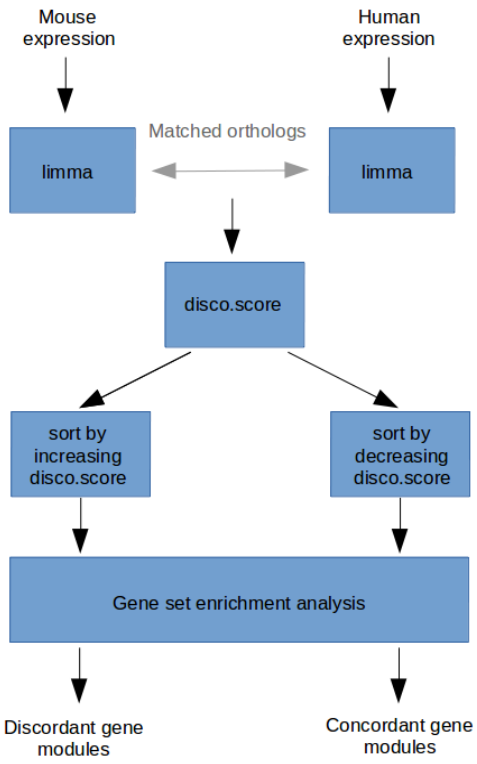


Figure 39 Algorithm used to identify concordant and discordant gene modules

The log fold changes and p-values between groups were calculated with R-package limma. The orthologous genes or genes corresponding to each other (if compared datasets derive from two groups of the same species) were mapped to each other. Then, disco.score was calculated for each pair of corresponding genes. GSE analysis was performed on the list of genes sorted by increasing disco.score to distinguish discordant gene modules and on the list of genes sorted by decreasing disco.score to distinguish concordant gene modules.

At the same time, I tested the performance of disco.score with $\log_2$FC values substituted by t-statistic for differential expression. In this case, the modified version of disco.score was illustrated by the formula:

$$disco.score = t.stat_{Hs} \cdot t.stat_{Mm}$$

Equation 6

where:

$t.stat_{Hs}$ - t-statistic for a gene in the human dataset, as calculated in differential expression analysis

$t.stat_{Mm}$ - t-statistic for a gene in the murine dataset, as calculated in differential expression analysis

The results of both disco.score methods were similar (Figure 40). Still, its first version remains more universal since depending on the method used to calculate differential expression the t-statistic is not always available, opposite to the universally used measure of gene expression change: log_2FC and p-values.

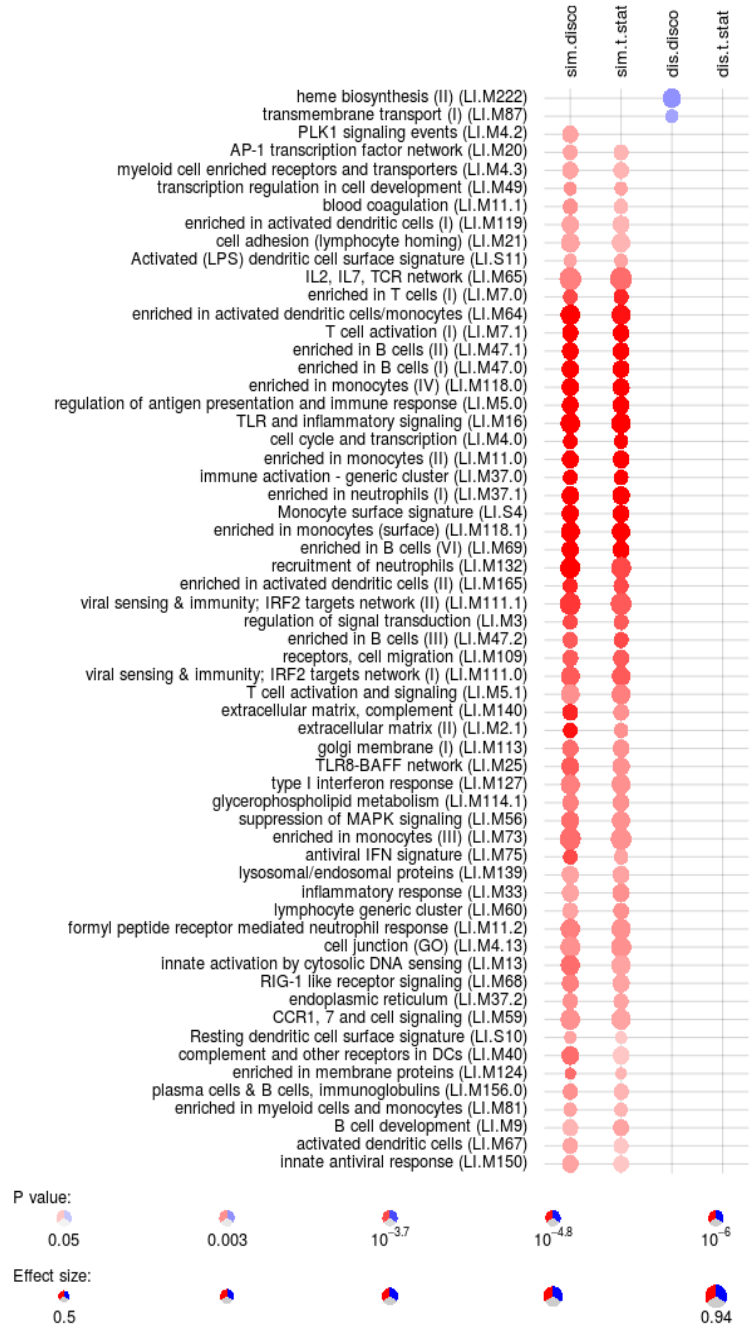


Figure 40 Sorting genes by disco.score results in more sensitive concordance and discordance detection compared with t-statistic

Concordant (red) and discordant (blue) modules enriched in comparison of WB expression profiles of patients from Gambia and 129S2 mice at day 21 p.i. detected with disco.score and on the basis of t-statistic. Only three modules vary between the results obtained using the two methods. P-value is illustrated by the intensity of the color and the effect size by the size of the dot. Only the modules with p-value for the enrichment lower than 0.005 are shown. The modules are described by the titles followed by the original number of genes in module and ID.

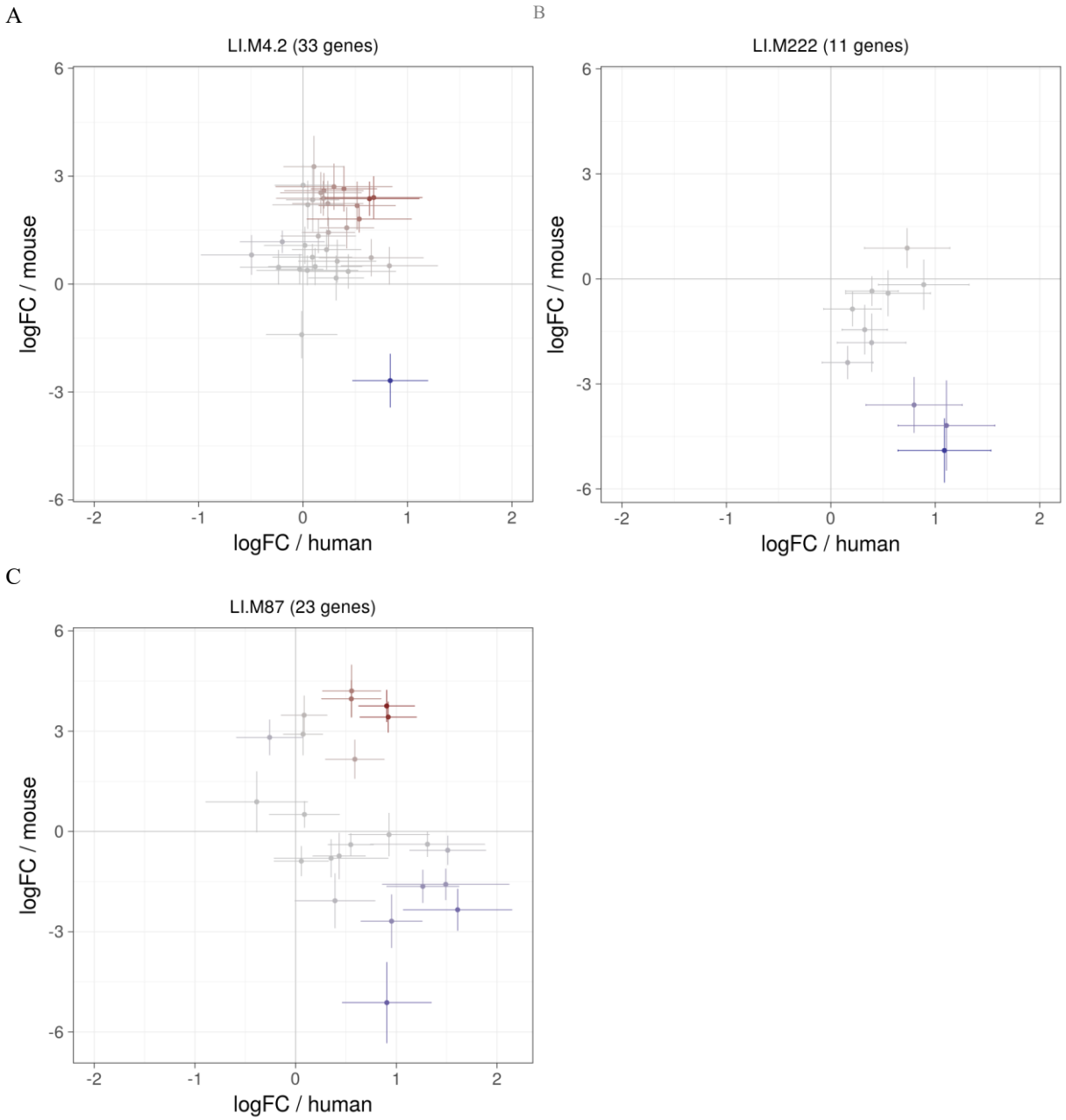


Figure 41 Three modules varying in the results obtained by disco.score and t-statistic

The classification of module LI.M4.2 as concordant as well as classification of modules LI.M222 and LI.M87 as discordant was detected by disco.score, but not by t-statistic gene sorting.

4.6. VALIDATION TESTS

4.6.1. Validation with simulated modules

Disco.score is a novel method to identify concordant and discordant elements of immune response in heterogeneous datasets, and, to my knowledge, the only such method dedicated not only to assess similarity between datasets but also to indicate their most dissimilar elements. Therefore evaluation of its performance is difficult due to lack of an objective measure of success. For this reason I validated the method in two contexts: first, to prove its technically correct performance, by simulating the concordant and discordant modules in two of the acquired datasets and detecting them with use of disco.score; and second, to verify the biological interpretability of the results, by identification of concordance and discordance in the datasets with known similarities.

4.6.2. Validation using two diseases with very similar transcriptomic profile

In the first step, I used the 129S2 mouse WB dataset and the human WB dataset from The Gambia (Maertzdorf, Ota, et al., 2011) to validate the ability of disco.score for detecting simulated concordant and discordant modules. I simulated a set of 100 concordant, 100 discordant and 100 random modules for every combination of two parameters: number of genes in a module (ranging from 10 to 50) and percentage of significantly regulated genes in a module (ranging from 10 to 30%). Using such defined modules I tested the ability of disco.score to correctly identify the concordant and discordant ones. The disco.score algorithm was able to correctly detect and classify the discordant and concordant modules and the sensitivity of detection increased with the larger number of genes per module and with the higher percentage of regulated genes per module. Not only the modules containing the discordant or concordant genes, but also the modules containing a mixture of discordant and concordant genes were detected correctly.

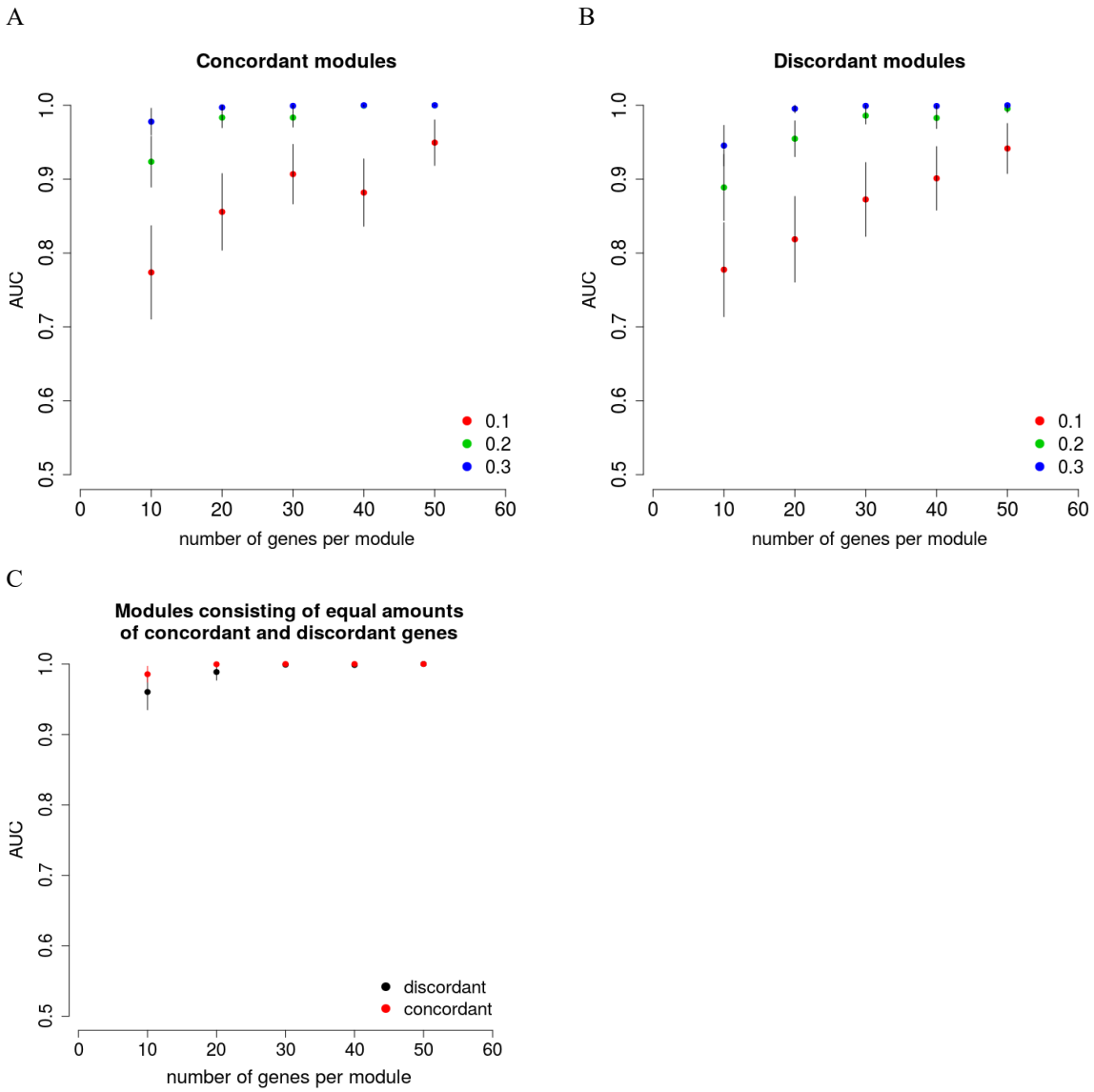


Figure 42 Results of the simulation test

Accuracy of the detection of **(A)** concordant modules, **(B)** discordant modules, and **(C)** modules concordant and discordant at the same, time illustrated by AUC corresponding to different numbers of genes in the modules and different percentage of regulated genes in the modules.

4.6.3. Validation using two cohorts of patients suffering of TB

In the next step, I tested whether the known biological similarities in two datasets are reproduced by disco.score. The transcriptomic regulation in patients suffering from TB and from sarcoidosis is highly similar compared to healthy individuals (Maertzdorf et al., 2012). In line with this observation, analysis of genes in Kyoto Encyclopedia of Genes and Genomes pathways (KEGG; (M Kanehisa & Goto, 2000; Minoru Kanehisa et al., 2017; Minoru Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016) revealed similar differential expression patterns in TB and sarcoidosis, including genes involved in systemic lupus erythematosus, complement and coagulation cascades, toll-like receptor signaling, and FcGR–mediated phagocytosis (Maertzdorf et al., 2012). Having acquired the WB transcriptional datasets from patients of both diseases I used disco score and GSE to identify

concordance and discordance between gene expression regulation in the two diseases. 83% of genes had a positive disco.score and the identified concordant gene sets were virtually identical as the gene sets enriched in comparison of both TB as well as sarcoidosis patients to healthy patients (Figure 43). There were no discordant modules enriched in the comparison of TB to sarcoidosis patients, which is in accordance with the previously described observation of lack of significant transcriptional differences in the patients of both diseases.

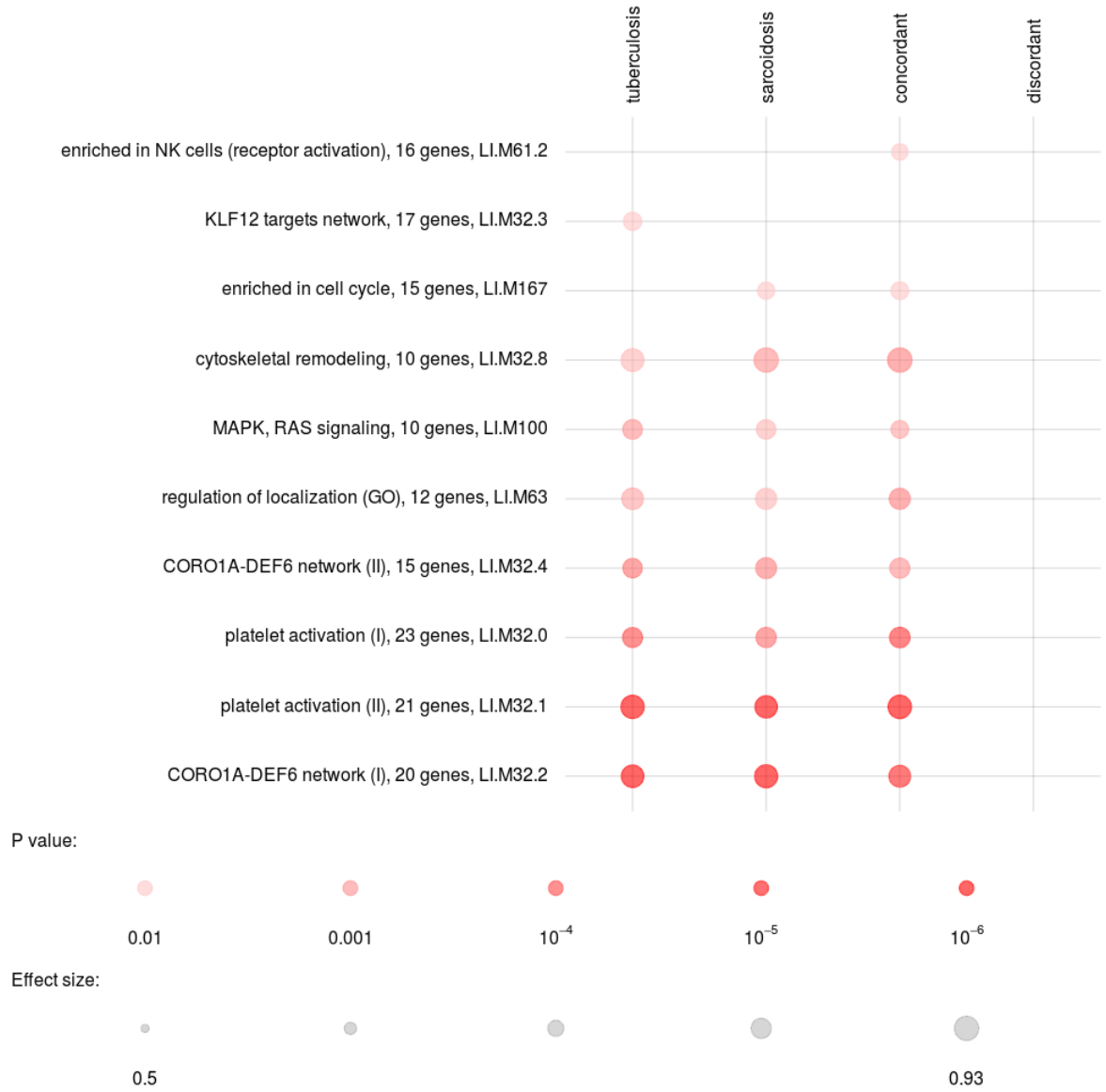


Figure 43 Disco.score-based concordance detection illustrates known biological background of disease similarity

P-value is illustrated by the intensity of the color and the effect size by the size of the dot. Modules enriched in test datasets derived from GEO (Maertzdorf et al., 2012, GSE34608). The gene modules enriched in TB patients, sarcoidosis patients, concordant gene modules identified with disco.score among the two groups of patients and discordant gene modules identified with disco.score are presented in the picture.

I performed another validation of disco.score by identifying similarities and differences in gene expression regulation among TB patients from Malawi and SA (Kaforou et al., 2013). Similarly as in case of TB and sarcoidosis patients, the differences between those two patient populations in

comparison to healthy were expected to be minor, and the similarities were expected to include the modules enriched in the two datasets in comparison to healthy. Visualization of the log_2FC values of the Malawian and the South African cohort plotted against each other and coloured by disco.score showed that the majority of the genes were regulated concordantly and were characterized by a positive value of disco.score (Figure 44).

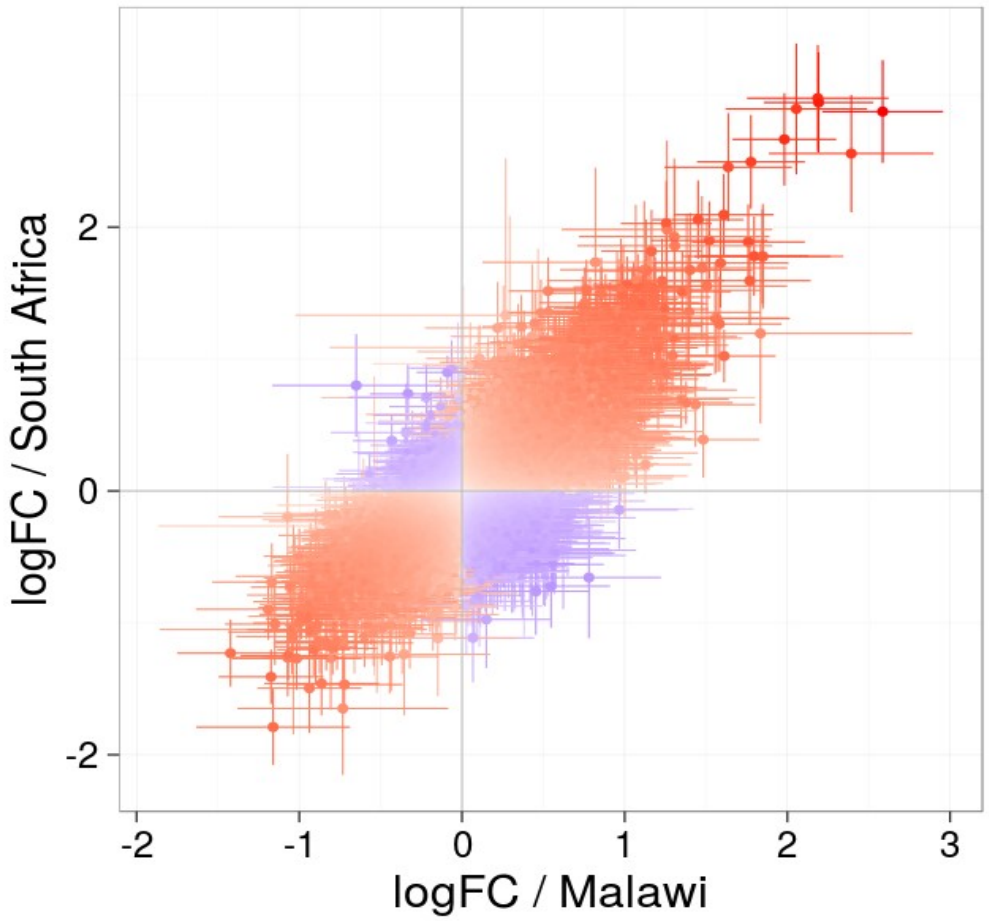


Figure 44 Distribution of disco.score in the assessment of similarity of gene expression changes in TB in a cohort from Malawi and cohort from SA

Increasing intensity of the red color indicates increase in disco.score and illustrates higher degree of similarity between gene expression in the two patient cohorts. Increasing intensity of the blue color indicates decrease in negative disco.score and a higher degree of dissimilarity in gene expression between two patient cohorts.

The GSE performed on the list of genes sorted by decreasing disco.score resulted in the identification of 94 concordant modules which were also enriched in the two datasets separately (Figure 45). The enriched modules contained elements characteristic for the TB response: T cell activation, DC signature, NK cell enrichment and IFN signaling. Sorting the genes according to increasing disco.score resulted in identification of only 4 discordant modules, related to transcription and to cell cycle.

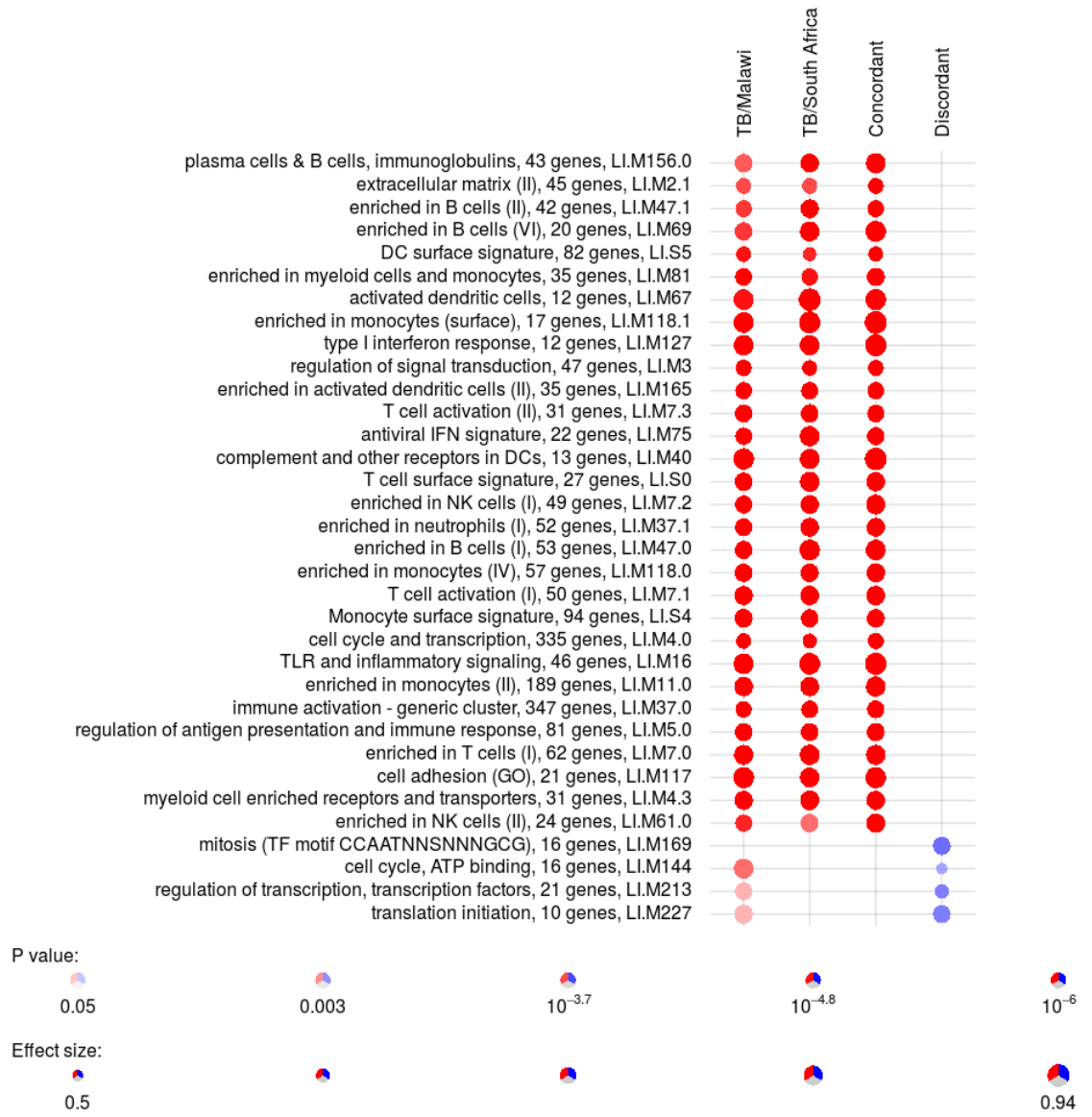


Figure 45 Modules enriched in test datasets from Malawi and SA

The gene modules enriched in TB patients from Malawi (TB/Malawi), TB patients from SA (TB/SA), concordant (red) and discordant (blue) gene modules identified with disco.score in the two groups of patients are presented in the picture. The modules are described by the titles followed by the original number of genes in module and ID.

I investigated the expression regulation of genes in each concordant and discordant module. The regulation of genes in the detected discordant modules was minute. The module "LI.M144 Cell cycle, ATP binding" identified as discordant possessed both discordant and concordant genes (Figure 46). In this case, the disco.score was sensitive to detect the discordant genes but failed to detect the concordant ones.

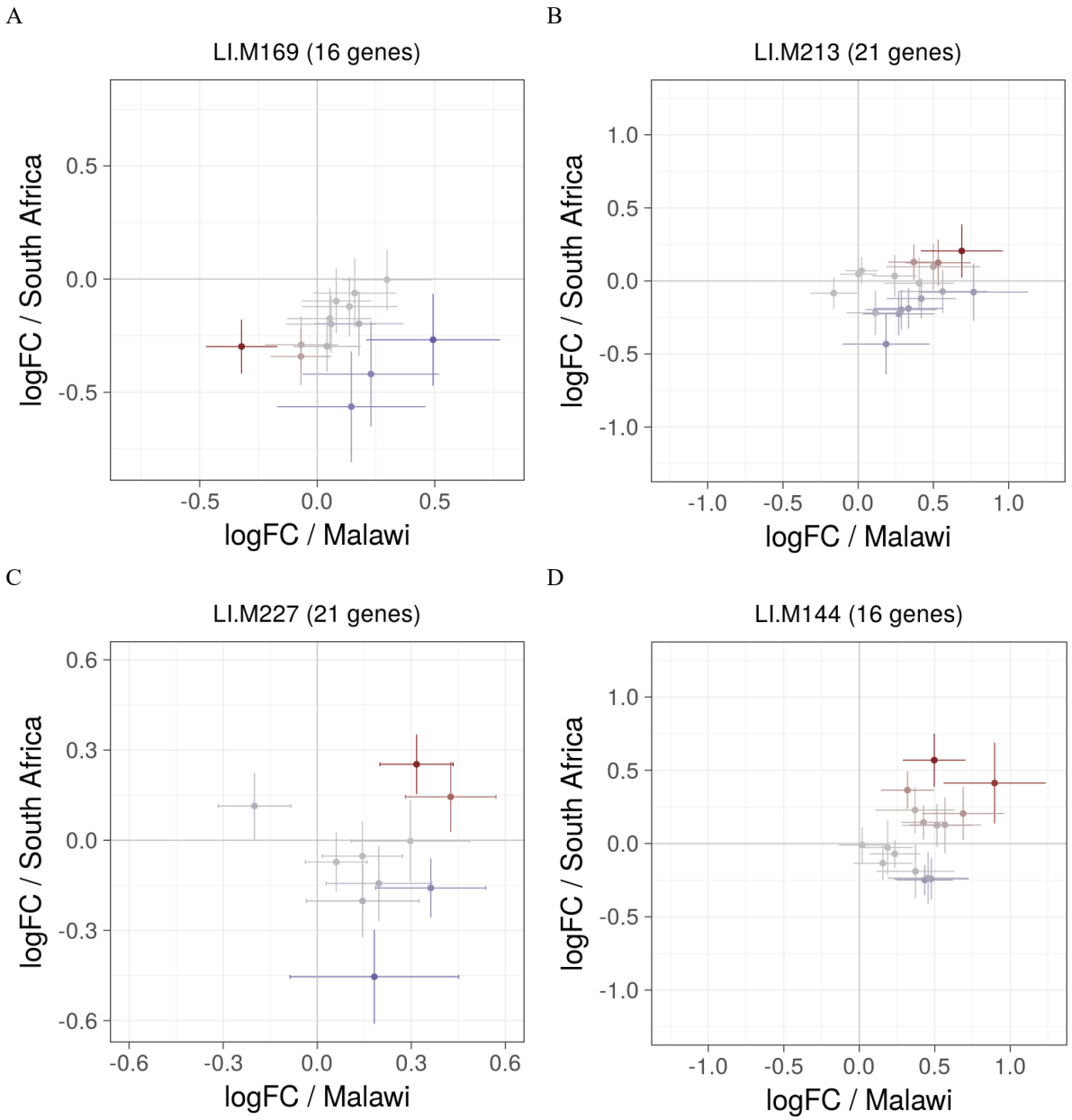


Figure 46 The modules assigned as discordant in the comparison of the South African and Malawian cohort

In the modules LI.M213 and LI.M144 concordantly regulated genes are also present.

The concordant modules consisted of genes regulated in the same directions and with similar magnitude (Figure 47).

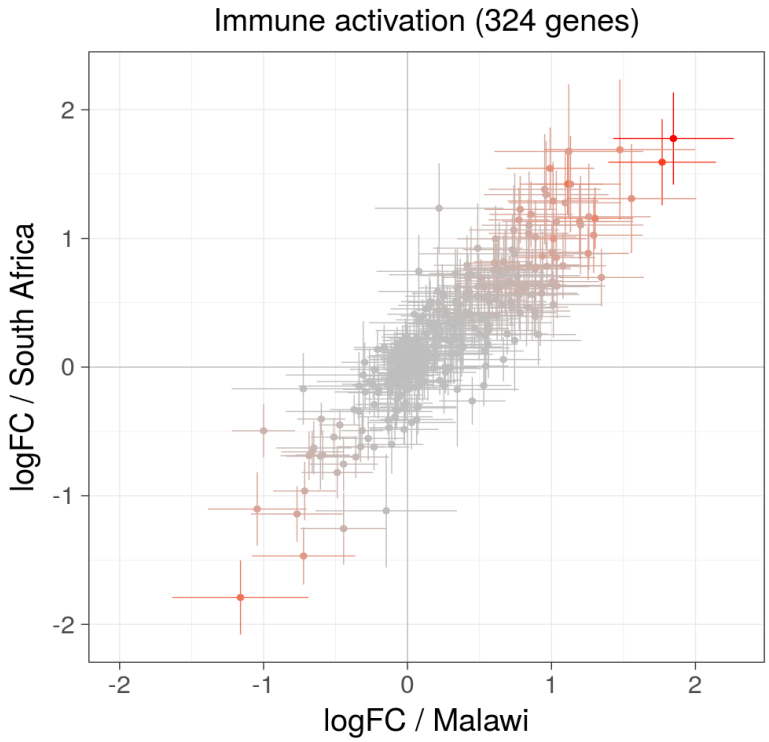


Figure 47 Log_2FC of gene expression of the cohort from SA plotted against log_2FC of gene expression of the cohort from Malawi

The plot presents the genes belonging to module "Immune activation - generic cluster", which was identified as concordant. The intensity of the color represents disco.score. Bars represent 95% confidence intervals (CI) for the log fold change.

4.6.4. Validation on human burn dataset and the corresponding mouse model

In the previously described studies approaching comparison of murine and human transcriptome data with correlation (Seok et al., 2013; Takao & Miyakawa, 2014) one of the investigated reactions was a response to burn trauma in patients as well as in C57BL/6J mice. Seok et al. (2013) indicated squared correlation coefficients between the datasets equal to 0.08 and identified "FcGR-mediated Phagocytosis in Macrophage and Monocytes", "IL-10 Signaling", "Integrin Signaling", "B cell receptor signaling" and "Toll-like receptor signaling" as the five most activated pathways in human burn. The r^2 for the correlation of the five most regulated pathways between man and mouse was ranging from 0 to around 0.5. Takao & Miyakawa (Takao & Miyakawa, 2014) excluded all the genes with $log_2FC < 2$ for man and $log_2FC < 1.2$ for mouse which resulted in the correlation coefficient between the same datasets equal to 0.68. They identified signaling pathways in which human and murine genes were regulated in the same direction, which included "Innate immune response", "Genes involved in Cytokine Signaling in Immune System" and "Lymphocyte Differentiation". I compared these results with the results of the disco.score-based concordance detection. I assigned the orthologs between human and murine genes from both datasets used by Seok et al. (2013) and Takao et al. (2014), separately in early response (time points up to 24 h after the burn trauma) and late response (time points between 24 h and 168 h after the burn trauma), calculated the disco.score for the orthologs

and performed GSE on the lists of genes sorted by disco.score. I identified concordant and discordant elements of immune response between C57BL/6J mice and man in early and late time points following the burn trauma (Figure 48, for visualization purposes only 35 modules are shown). There were 68 concordant modules between mouse and man in the first day after infection. They included several modules related to NK cells and to innate immunity, e.g. the module "immune activation – generic cluster" (LI.M37.0) which was most significantly enriched, modules related to adaptive immunity, like "T cell activation" or the module "antigen presentation (lipids and proteins)" (LI.M28) which had the largest effect size (Figure 48). Additionally, there were two discordant modules: "NK cell surface signature" (LI.S1) and a non-annotated module (LI.M151, not shown), however the enrichment in the discordant modules was characterized by larger p-value and smaller effect size than in the concordant modules. One week after stimulation many of the adaptive immunity-related modules were still regulated in the human dataset, but not in the murine one. At that time point the concordances encompassed innate immunity and metabolism. The two discordant modules detected at this time point were "type I IFN response" (LI.M127) and "NK cell surface signature" (LI.S1).

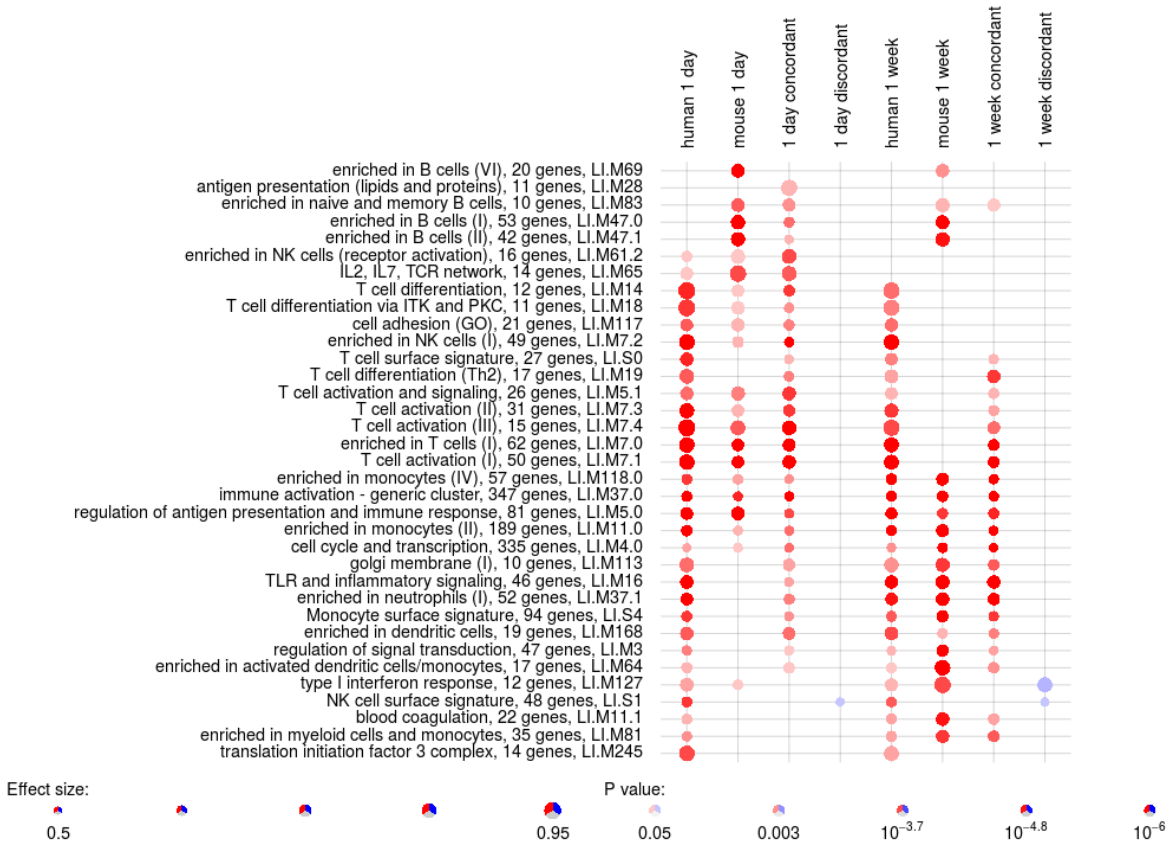


Figure 48 Concordant and discordant modules enriched in burn datasets

The datasets were derived from GEO (Calvano et al., 2005; GSE3284). The gene modules enriched in patients after burn and mouse model of burn in time points of 1 day and 1 week are presented in the picture. P-value in illustrated by the intensity of the color and the effect size by the size of the dot. Only the modules with p-value for the enrichment smaller than 10^{-7} are shown. The modules are described by the titles followed by the original number of genes in module and ID.

The previous studies (Seok et al., 2013; Takao & Miyakawa, 2014) identified the innate immunity-related modules as similar in the human and murine response to burn, which is compatible with the results obtained using the disco.score even though the calculated correlation coefficients between murine and human expression values are low. In contrast to the preceding studies, which focused entirely on the detection of similarities, the application of disco.score enabled as well identification of the opposite gene expression change in NK cells or IFN modules (Figure 49). Briefly, disco.score algorithm-based analysis not only identified previously described similarities between human and murine burn datasets, but also indicated gene modules regulated in opposite manner between the two species.

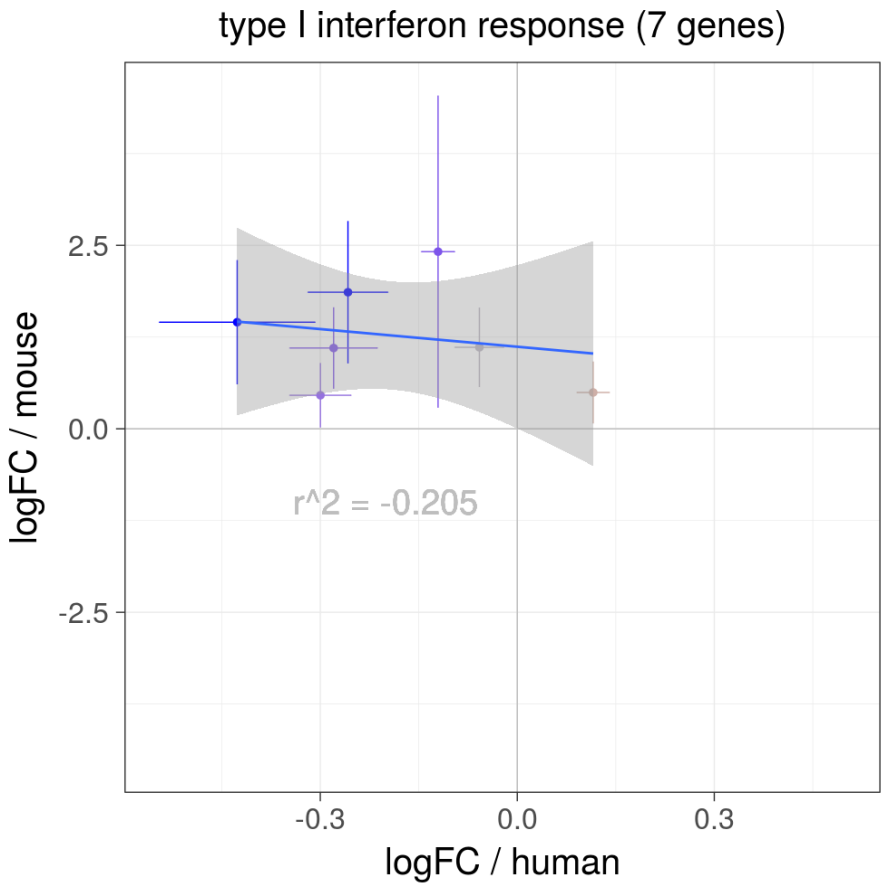


Figure 49 The module "Type I IFN response" is discordant one week after the burn

4.7. DISCO.SCORE IDENTIFIES CONCORDANCE AND DISCORDANCE OF RELATED HUMAN AND MURINE DATASETS IN TB

I acquired publicly available datasets from TB patients and murine TB. My colleagues from MPIIB Department of Immunology, Lisa Scheuermann and Anca Dorhoi, conducted the experiments to acquire blood from two infected and uninfected mouse models of TB, the low susceptible C57BL/6 mouse strain and the highly susceptible 129S2 mouse strain as well as from Mtb stimulated and unstimulated human macrophage-like THP1 cells. Karin Hahnke from MPIIB, Department of Immunology and Hans J. Mollenkopf from Microarray Core Facility, MPIIB prepared the samples and conducted microarray experiments on the acquired tissues. For each acquired dataset I calculated differential expression of genes between infected and uninfected groups. Next, I calculated the disco.score for each pair of orthologous genes and identified concordant and discordant immune modules and MSigDB Hallmark Gene Sets (Figure 50) and verified concordance of gene expression change by visualizing the gene expression in the identified concordant and discordant modules. The genes belonging to concordant modules were regulated in the same direction and the majority of them (for example 56% in comparison 2) had non-negative weighted correlation coefficients. In contrast to the assignment of genes which were significantly regulated in the same direction in both mouse and man but at the same time characterized by negative correlation or correlation coefficient close to 0 as ‘not similar’ by the correlation approach, such modules were identified as concordant by the disco.score approach. Similarly, modules containing genes regulated in opposite direction were identified as discordant even if they presented positive correlation coefficient. If not indicated otherwise, the results presented in the following text are based on transcriptomic modules created by Li et al. (2014).

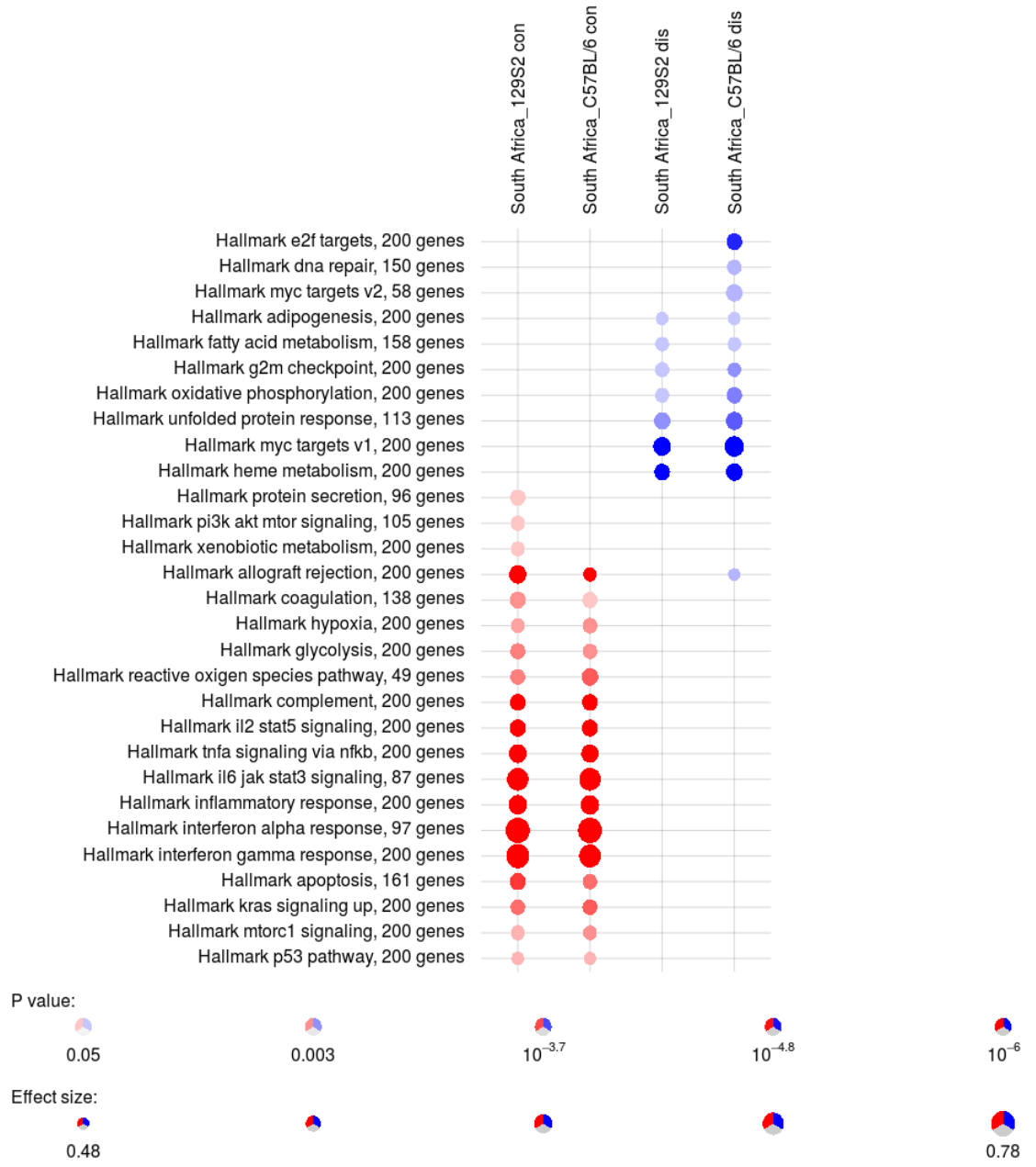


Figure 50 Results of disco.score based module detection with use of MSigDB modules in comparison of human and murine datasets

Concordant (red) and discordant (blue) MSigDB Hallmark Gene Sets enriched in human WB dataset from SA and WB datasets from C57BL/6 and 129S2 mice at day 21 p.i. The modules "Hallmark IL2 STAT5 signaling" and "Hallmark IL6 STAT3 signaling" are concordant in comparison of South African cohort to both 129S2 and C57BL/6 strains. The modules are described by the titles followed by the original number of genes in module and ID.

Concordant modules were present between both mouse strains and man at every time point p.i. of mice and their number increased towards day 21 p.i. of mice (Figure 51 and Figure 52). The number of discordant modules, however, was highly dependent on the compared mouse strain: it decreased with time p.i. in the comparison of man and 129S2 mice, but remained at high level independent of the time point in the comparison of man and C57BL/6 mice (Figure 53 and Figure 54).

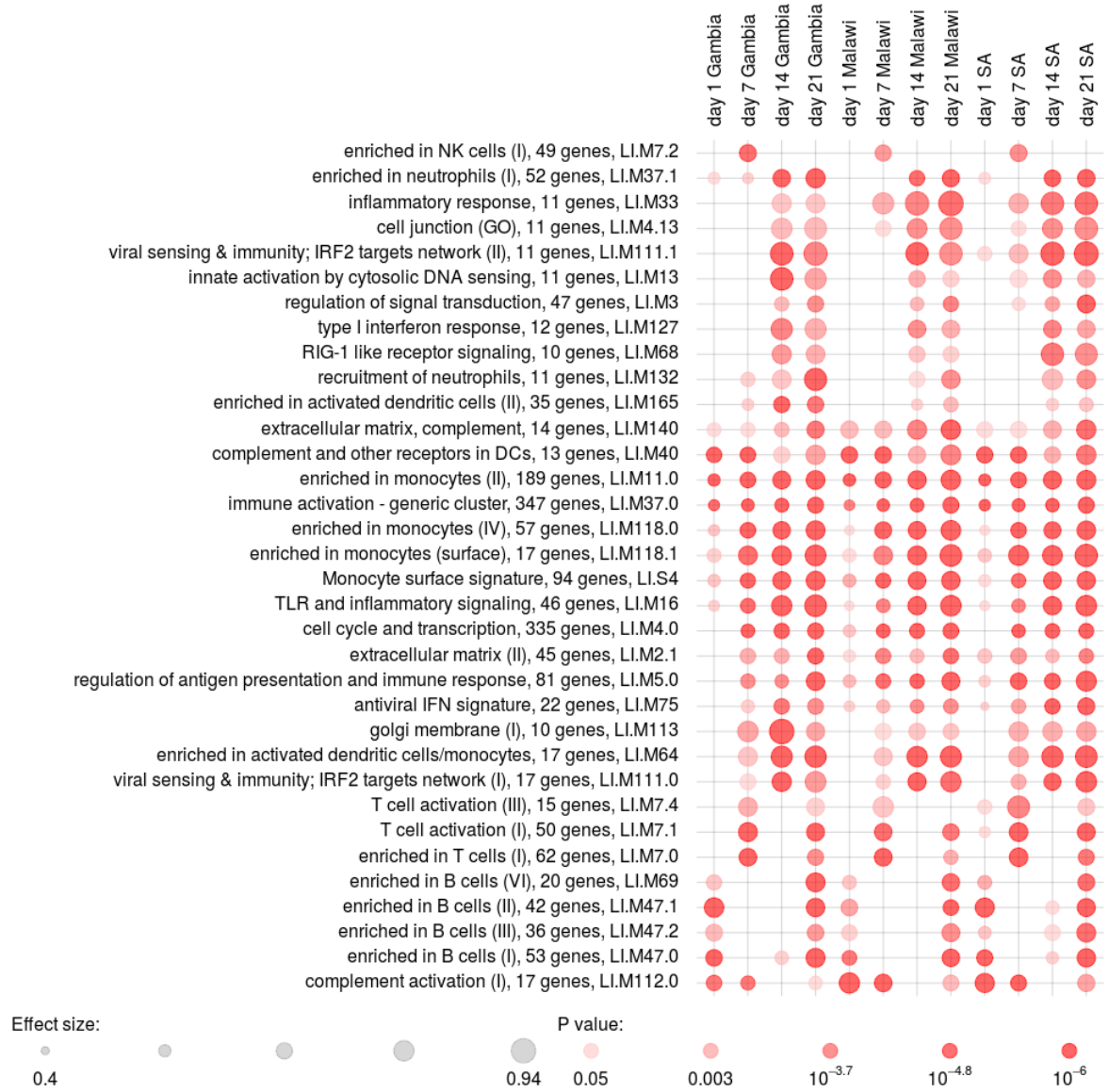


Figure 51 Concordant modules in comparisons of 129S2 WB with human datasets

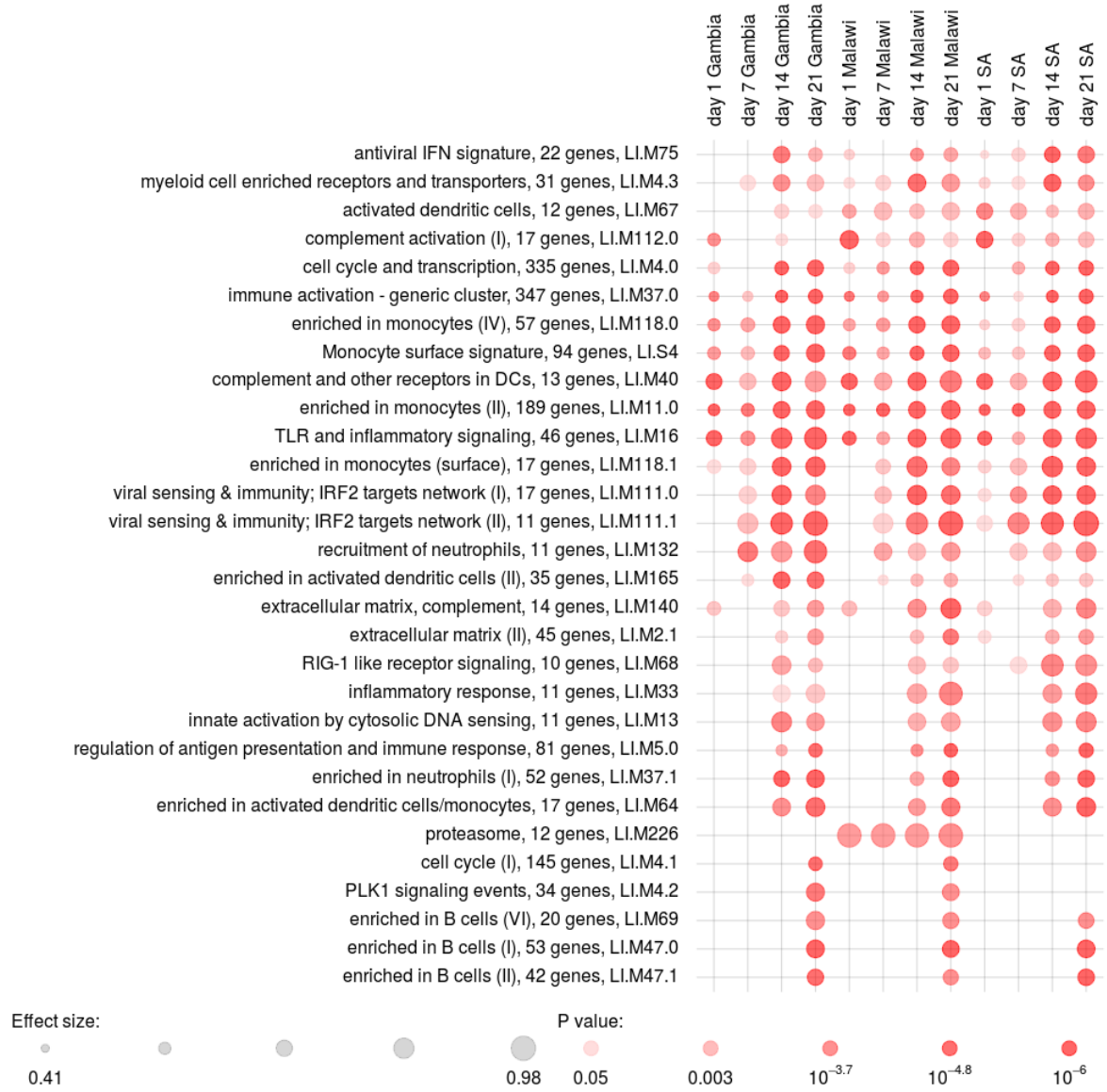


Figure 52 Concordant modules in comparisons of C57BL/6 WB with human datasets

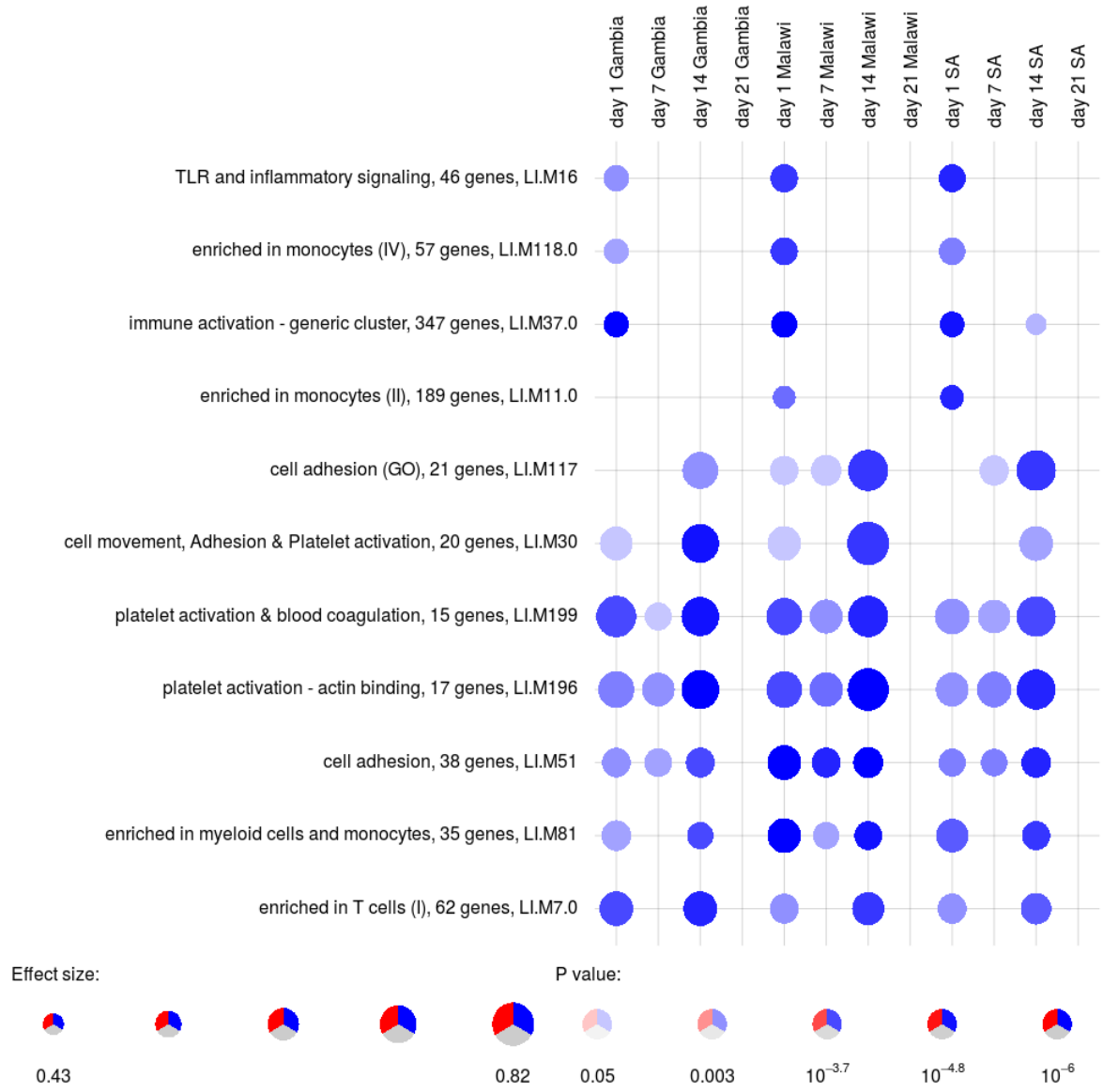


Figure 53 Discordant modules in comparisons of 129S2 WB from different time points with human datasets

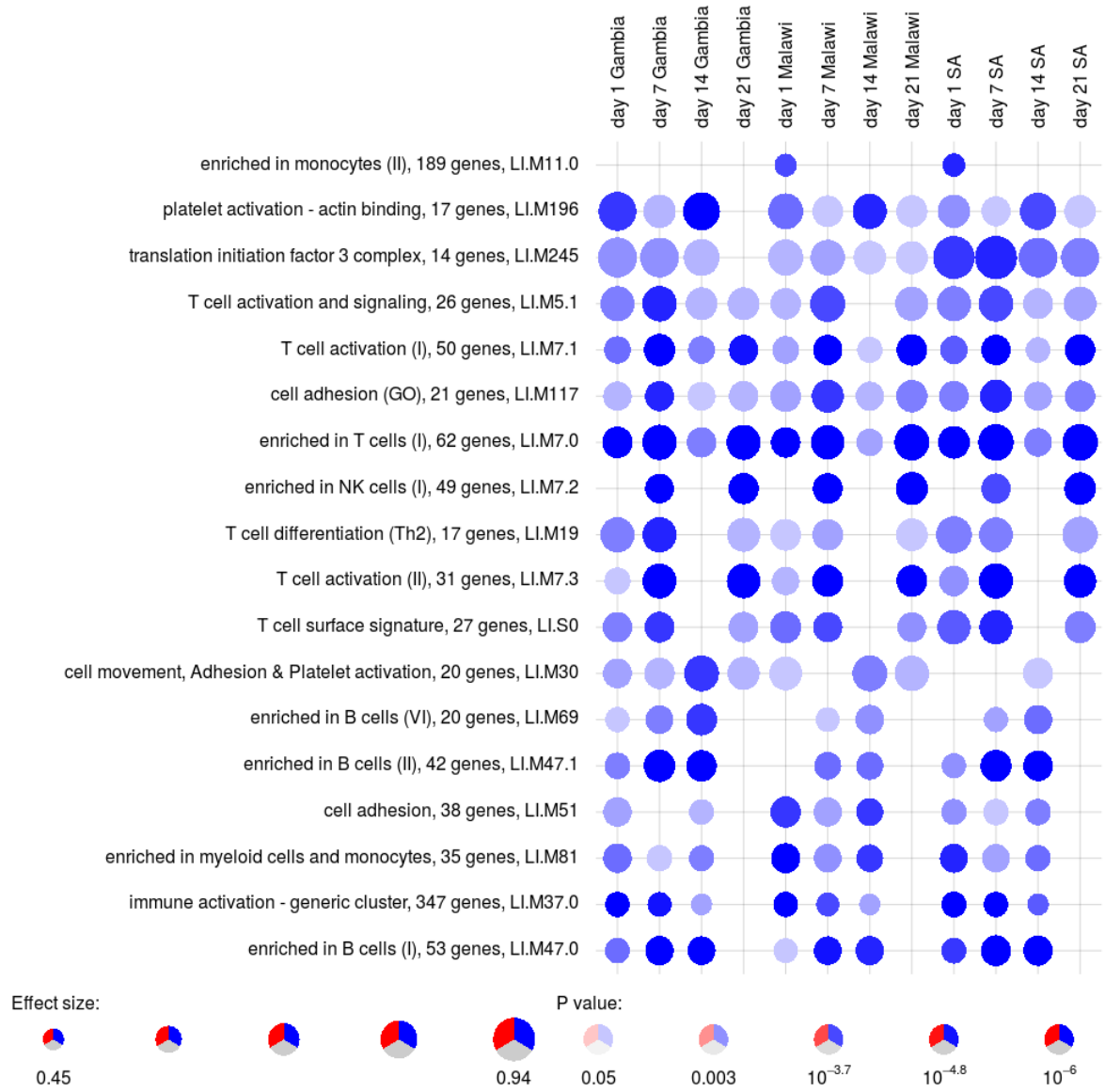


Figure 54 Discordant modules in comparisons of C57BL/6 WB from different time points with human datasets

4.8. SIMILARITY OF MURINE AND HUMAN RESPONSES TO INFECTION CHANGES OVER TIME

The blood samples analyzed by microarrays from 129S2 and C57BL/6 mouse strains were collected before infection (day 0) as well as at the following time points: day 1, 7, 14 and 21 p.i.. I calculated differential gene expression comparing each time point p.i. with the day 0 as control healthy sample. Next, I compared the time series data from both mouse strains with publicly available datasets from human cohorts from The Gambia (Maertzdorf, Ota, et al., 2011), SA and Malawi (Kaforou et al., 2013), which included HIV negative (HIV-) TB patients, and HIV- and HIV+ individuals with latent TB infection (LTBI) from the same locations as controls. In every comparison of each human dataset with each time point p.i. in both mouse strains I detected concordant and discordant gene modules. Similarly, I compared the corresponding time points p.i. between the two mouse strains to test whether the phenotype differences in their reaction to Mtb infection are illustrated by concordant and discordant gene expression. For both highly susceptible 129S2 and low susceptible C57BL/6 mouse strain the amount of significantly differentially regulated genes as well as the degree of concordance with any of the human datasets increased towards day 21 p.i (Figure 55), reaching above 100 concordant modules for the 129S2 and above 70 for the C57BL/6 mouse. The highest number of 29 discordant modules between the human and murine data appeared on day 1 p.i. of mice. In such an early time point, the immune response to TB is not fully established yet and is referred to as early response. Notably, there was a different trend in the amount of the identified discordances between the two mouse strains and man: while their number decreased towards day 21 p.i. in the comparison of human data with 129S2 mouse data, it remained at a high level of around 40 modules in the comparison of human vs C57BL/6 strain. In comparison of the two mouse strains the number of concordant modules doubled between day 1 and day 14, but only sparingly increased between days 14 and 21 p.i.. There were no discordant gene modules in comparison of the two mouse strains up to day 14 p.i., which changed in the day 21 p.i. when 24 modules containing genes regulated in opposite directions were observed.

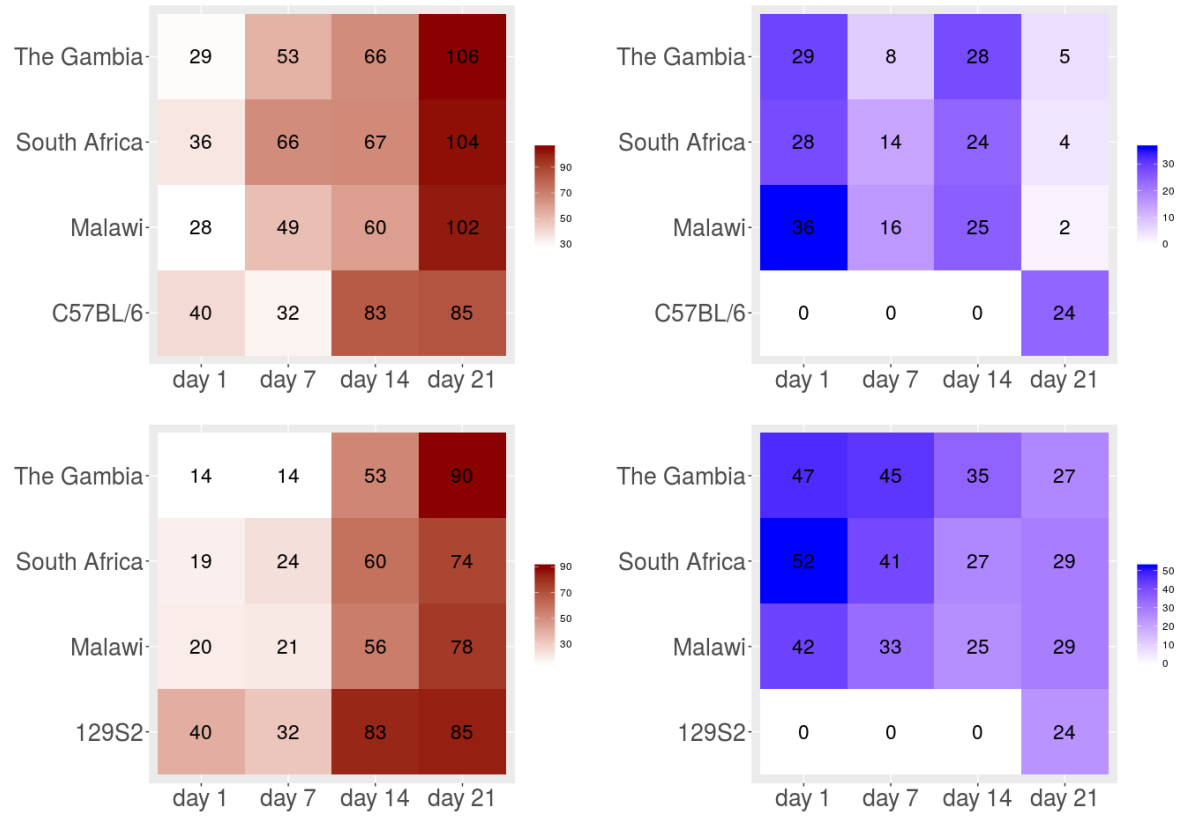


Figure 55 Module counts in comparisons of different human and mouse datasets

Red color refers to concordant and blue to discordant modules. Upper panel: Comparisons of 129S2 mouse strain data with human datasets from cohorts from The Gambia, SA, Malawi and with C57BL/6 mouse strain. Lower panel: Comparisons of C57BL/6 mouse strain data with human datasets from cohorts from Gambia, SA, Malawi and with 129S2 mouse strain.

4.9. DISCORDANCE IN 129S2 AND C57BL/6 GENE EXPRESSION CHANGES CORRESPONDS WITH THE HIGHLY SUSCEPTIBLE PHENOTYPE

Different susceptibility to low dose aerosol Mtb infection characterizes the 129S2 and C57BL/6 mouse strains. The highly susceptible 129S2 mice suffer from progressive TB and succumb to disease within 40 days p.i., while low susceptible C57BL/6 mice develop chronic TB and survive for more than 100 days p.i.. At very early time points p.i. there are no visible phenotypic differences in the disease development between C57BL/6 and 129S2 mouse strains. In the days 7-10 p.i. inflammatory cells such as neutrophils begin to infiltrate the site of infection in 129S2 mice, which is not observed in the C57BL/6 mice until approximately day 14 p.i. By 21 day p.i. the 129S2 strain develops severe lung pathology characterized by large, necrotic lesions, whereas in the C57BL/6 strain smaller non-necrotic lesions with less inflammation are formed.

I expected those differences to be illustrated on the transcriptional level. To test this hypothesis, at each investigated time point p.i. I compared the gene expression in the two mouse strains. There were no discordances present on the day 1 p.i. and the concordant modules encompassed "Complement and other receptors in DCs" (LI.M40), "complement activation" (LI.M112.0), "enriched in monocytes" (LI.M11.0), "immune activation- generic cluster" (LI.M37.0). All the enriched concordant modules were related to innate immunity, immune signaling and platelets. The same similarities remained present on day 7 p.i., when also spliceosome and proteasome related genes became co-regulated. On day 14, the amount of concordantly regulated modules further increased with the TB characteristic modules - "type I IFN response" and "antiviral IFN signaling" added to the ones present in earlier time points. The same set of modules additionally complemented by cell cycle related genes remained concordant on the day 21 p.i..

At this last investigated time point some sudden differences appeared between the two mouse strains. A set of 13 T-cell related modules was identified as discordant, including "T cell activation and signaling" (LI.M5.1), "enriched in T cells" (LI.M7.0) and "enriched in NK cells" (LI.M7.2). Those modules were also identified as discordant between C57BL/6 mice and men. For the first time, the number of concordant modules for the two mouse strains was lower than the number of concordant modules between the 129S2 mice and man. Among the concordant modules between the two mouse strains there were modules related to innate immunity detected also at the earlier time points p.i.. In summary, the T-cell response has been identified as the major difference between 129S2 and C57BL/6 mice at day 21 p.i., when the disease progressed more profoundly in the highly susceptible compared to the low susceptible mouse strain.

4.10. T CELL CO-RECEPTOR GENES DRIVE THE DISCORDANCE BETWEEN HIGHLY SUSCEPTIBLE AND LOW SUSCEPTIBLE MICE

I further investigated the expression of genes in four of the modules identified as discordant between 129S2 and C57BL/6 mouse strains on day 21 p.i.: "enriched in T cells (I)" (LI.M7.0), "T cell activation and signaling" (LI.M5.1) and "T cell activation (I)" (LI.M7.1). Between day 14 and 21 p.i. those genes were regulated in opposite directions between the strains (Figure 56). The genes driving those differences included Cd28, Cd3d, Cd3e, Cd3g, Cd5, Cd7 and Cd96, which encode T-cell co-receptors. They were up-regulated in the low susceptible strain C57BL/6 but down-regulated in the highly susceptible 129S2 strain. To investigate how those genes are regulated in human TB patients in comparison to the both mouse strains I analyzed their expression in The Gambian cohort. Interestingly, the genes were highly concordantly regulated between the 129S2 mice and humans, and highly discordantly regulated between C57BL/6 mice and humans (Figure 57). Therefore, I identified 16 genes

related to T cell co-stimulation with opposite expression changes in human and highly susceptible 129S2 mice vs low susceptible C57BL/6 mice.

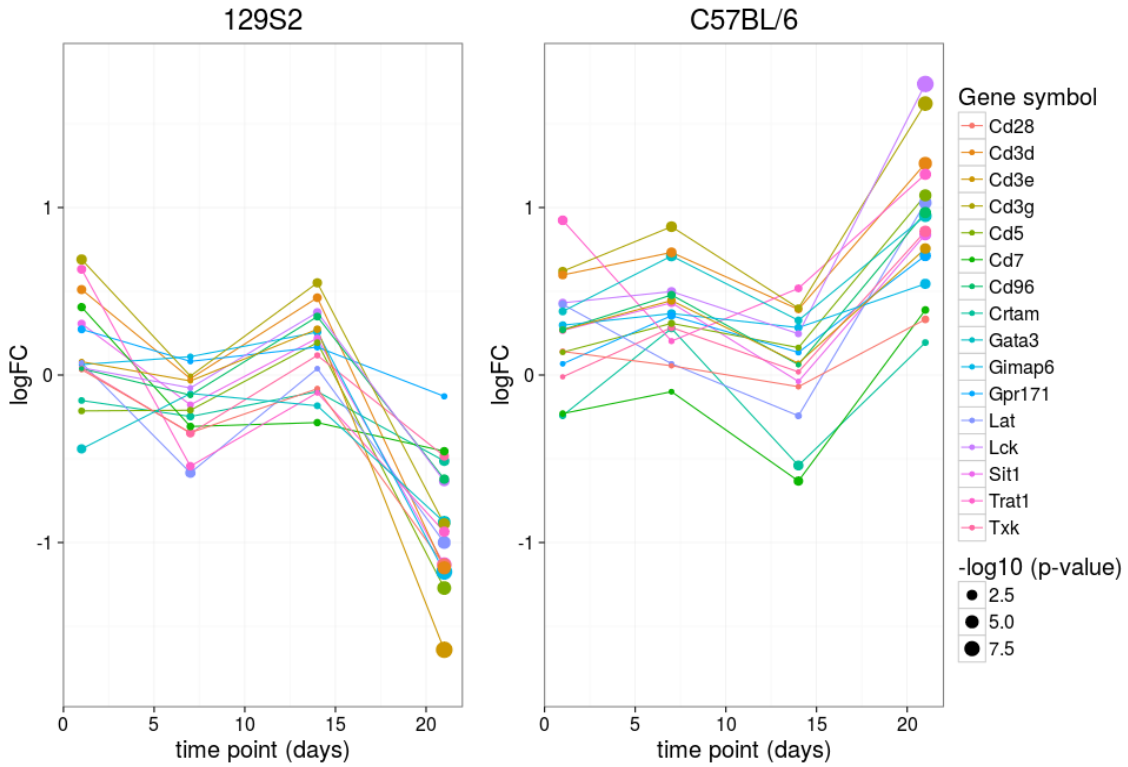


Figure 56 Expression changes of selected genes belonging to the T-cell related modules

The pictured genes belong to the modules: "enriched in T cells (I)", "T cell activation and signaling" and "T cell activation (I)". The illustrated time points are: day 1, day 7, day 14 and day 21 p.i.. The selected genes drive the differences in the patterns of T-cell expression changes in **(A)** 129S2 and **(B)** C57BL/6 mice.

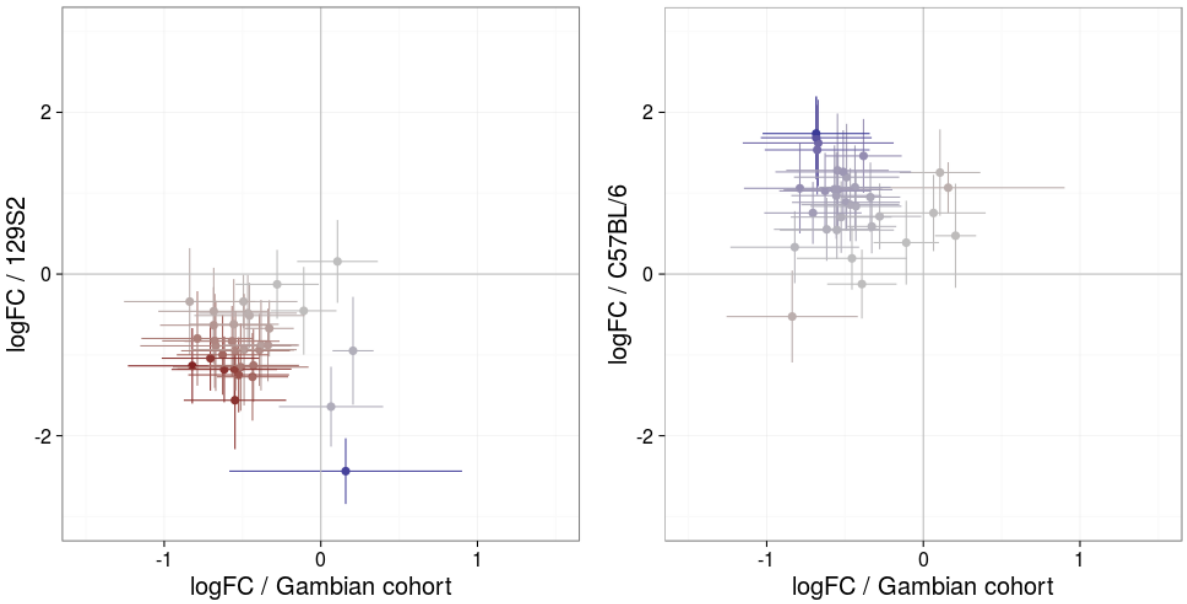


Figure 57 Log_2FC of the set of 16 genes plotted for mouse data vs data from patient cohort from Gambia

(A) 129S2 mouse strain data plotted vs human data; **(B)** C57BL/6 mouse strain data plotted vs human data. Bars represent 95% confidence intervals for the log fold change.

4.11. GENE EXPRESSION IN RESPONSE TO MTB INFECTION IS CONCORDANT IN HUMAN AND MURINE MACROPHAGES

The discrepancies between human and murine C57BL/6 response to TB were related entirely to the adaptive immunity. Since macrophages are a crucial subset of cells taking part in Mtb infection and at the same time are innate immune cells, I investigated whether the discordances found in blood are also present in gene expression profiles of macrophages. In the first step I calculated differential gene expression between human THP1 cells 6 h p.i. and uninfected cells in a dataset collected by my colleague from MPIIB, Department of Immunology, Anca Dorhoi as well as in a murine dataset derived from a study by McNab et al. (2013) (comparison 26). 34 concordant modules including "antiviral IFN signature" (LI.M75), "RIG-1 like receptor signaling" (LI.M68), "chemokine cluster (I)" (LI.M27.0) and no discordant modules were detected in this comparison. This observation was reproduced in the other datasets derived from macrophages (Carow et al., 2011; Kaforou et al., 2013; Thuong et al., 2008; comparisons 25, 27-32 and 34-36). Those datasets covered a broad range of TB cases and time points p.i.: for example, in the dataset collected by Thuong et al. (2008) the macrophages were derived from patients who had recovered from pulmonary TB (samples referred to as PTB), TB meningitis (samples referred to as TBM) and individuals with LTBI. The macrophages were derived from isolated PBMCs and infected with Mtb for 4 h. I compared the LTBI samples (comparison 27), PTB samples (comparison 29) and TBM samples (comparison 31) with mouse bone marrow derived macrophages (BMDMs; Carow et al., 2011) non-stimulated and infected for 24 h with Mtb.

22 modules were universally identified as concordant in each performed macrophage comparison (comparisons 25-36, Figure 58). They included the TB-characteristic IFN response related genes: "antiviral IFN signature" (LI.M75), "type I IFN response" (LI.M127), "chemokine clusters" (LI.M27.0, LI.M27.1), and "innate activation by cytosolic DNA sensing" (LI.M13), "cell cycle and growth arrest" (LI.M31) and "enriched in activated dendritic cells/monocytes" (LI.M64). No overlapping discordant modules were detected in the comparisons of different human and murine macrophage datasets. The detected concordances in the early time points after infection were related to innate immunity (e.g. "innate activation by cytosolic DNA sensing" (LI.M13), "chemokine cluster" (LI.M27.0), "chemokines and receptors" (LI.M38). The time points of 24 h p.i. presented also concordant IFN response and DC activation.

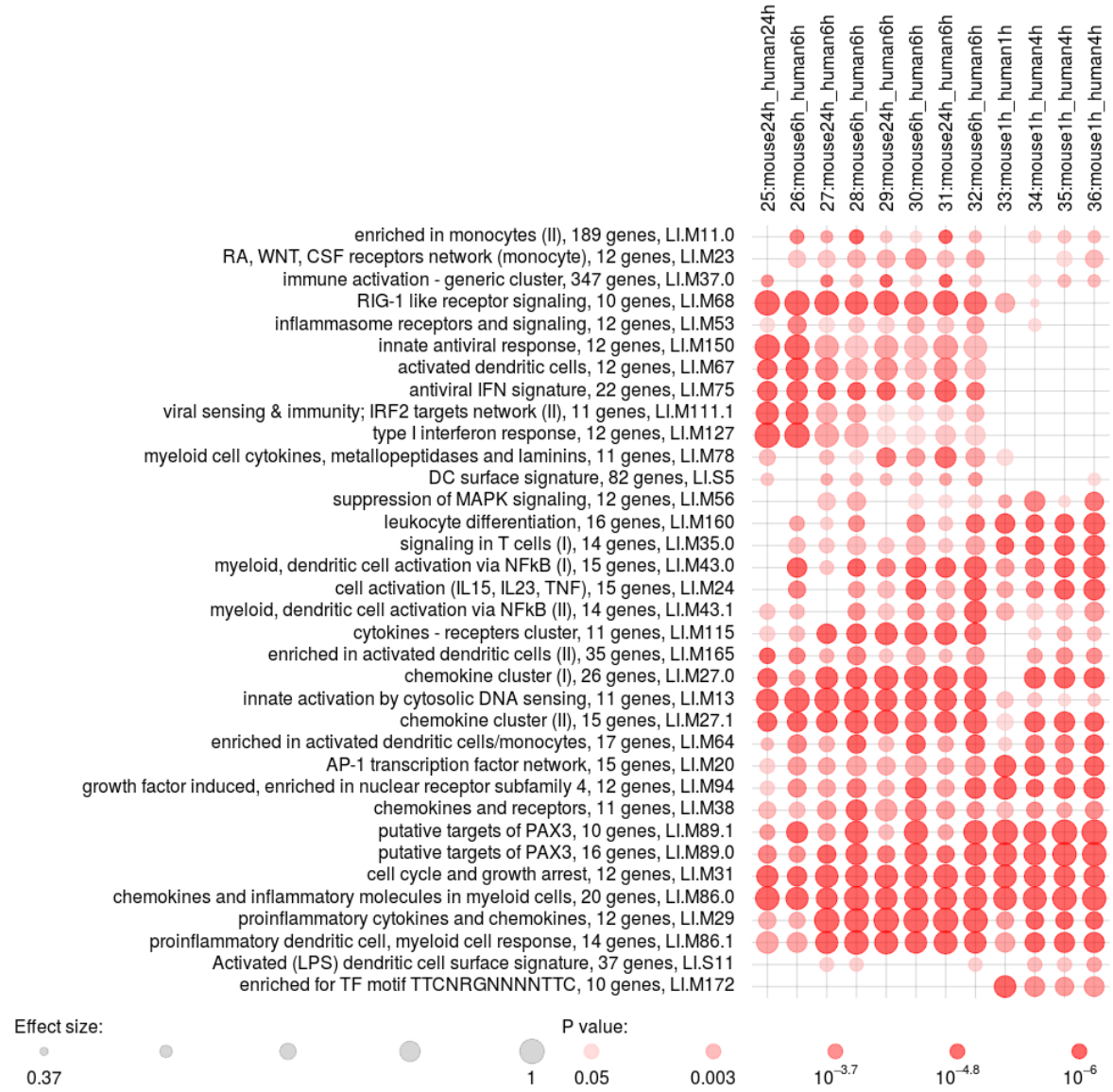


Figure 58 Concordant modules in the comparisons of murine and human macrophages

While the WB response to Mtb infection involved T- and B-cell signaling, in macrophages the enrichment was mostly clustered in cell cycle, metabolism and innate immunity. As stated before, macrophages are critical elements of the first line of defense against Mtb infection (Dorhoi & Kaufmann, 2015; Thuong et al., 2008). Absence of discordant modules indicates that the expression regulation of the macrophage response to Mtb is largely conserved in mouse and man.

5. CHAPTER 5: DISCUSSION AND CONCLUSIONS

In the last chapter of my thesis I summarize the conducted analysis and comment the obtained results. Advantages and flaws of the presented methods and approaches are addressed. I compare the findings described in this thesis to the previously published statements. I describe the conclusions derived from analysis of individual variability among TB patients in the first part of the chapter and those derived from the comparison of transcriptomic responses to TB in man and mouse in the second part of the chapter, and describe how the developed methods and pipelines could be used for the analyses of different problems.

No scientific work can be contained in one independent publication, because it is inspired by previously performed research or previously asked questions, and it should still inspire scientists to continue, validate or contradict it in the future. This thesis is deeply rooted in the questions arising around the problem of understanding TB as well as in everyday questions asked by the scientists in a TB laboratory, for example: which animal model to use to mimic certain aspect of TB? It also makes use of multiple previously published studies and data collections. In the discussion of my results I look at the results again from a wider perspective of the TB research field and suggest how will those results influence the current understanding of TB and how can future research benefit from the results presented in this thesis.

5.1. THE ACHIEVEMENTS OF THIS THESIS

5.1.1. Analysis of individual variability among TB patients

In the presented analysis of individual variability among TB patients I have shown that subgroups of them are characterized by various activation profiles of immune response. Specifically, I focused on the scale of IFN signaling in particular patients and have shown that there are TB patients who do not develop strong IFN responses when undergoing TB which is correlated with less severe disease.

In order to show it, I collected publicly available data from seven TB patient cohorts into one meta-dataset and performed GSEA for each patient on the list of genes sorted by the z-score corresponding to the probability that the expression value of particular gene belongs to the distribution of the expression values of this gene in healthy individuals. The results clearly showed that even though enrichment of modules including T-cells, B-cells, innate immunity, IFN signaling, monocytes and many others typically enriched in TB patients is significant in majority of the TB patients, in each study cohort there are individuals who lack this characteristic enrichment. Moreover, there were visible patterns of gene expression among the TB patients which were independent of the study they originally participated in. This raised further questions I asked in this thesis: are there several coherent, identifiable genetic profiles that can be activated by different individuals in response to TB? What are the correlates of the differences in transcriptomic profiles of the TB patients? To answer these questions, I focused on one of the most striking differences among the studied patients: IFN response, which has been previously described as dominant response in TB on transcriptomic level.

Division into IFN- and IFN+ patients

Two types of IFN signaling which have been shown to be crucial for the outcome of TB are (i) the mostly detrimental type I IFN signaling pathway and (ii) the protection mediated by IFN type II signaling pathway. However, around 17% of the TB patients collected in the training MDS did not present enrichment in the modules related to IFN signaling. Out of the patients with enrichment in the IFN modules, 89% were enriched in both IFN type I and II IFN modules, 6% in IFN type I and 5 % with IFN type II modules only. To further investigate the patients presenting enrichment in IFN type I gene modules, I split the cohort of TB patients from the MDS into IFN I positive and IFN I negative donors based on the enrichment in IFN type I modules and further investigated the differences between them.

Investigation into the differences in the gene expression between the IFN- and IFN+ patients

In the first step I used logistic regression models and unsupervised ML models to test whether the IFN status is related to any other known factors influencing the datasets: the HIV status of the patients, their ethnicity and residence, and the microarray platforms used to conduct experiments. TB was the most significant factor influencing the IFN+ status; on the other hand, coinfection with HIV and *Streptococcus sp.* and *Staphylococcus sp.* was also significant. GSEA on the weights of genes calculated in PCA in the components differentiating between IFN+ and IFN- patients has shown that contribution of T cells and NK cells is dominant in this division. PCA of samples from TB patients revealed that even though the data has been normalized, the differences between datasets derived from different studies still stratify the data. However, it does not translate into significant biological differences: the enrichment in genes sorted by weights in the principal components differentiating between the studies resulted in only 4 significantly enriched modules.

The differences in the enrichment of IFN related modules could be a consequence of varying levels of IFN-α, IFN-β or IFN-γ signaling molecules in the WB. However, the IFN+ and IFN- patients presented similar expression of IFNA2, IFNB1 and IFNG genes. Despite that, there were significant differences in the expression of IFN-α and IFN-γ receptor genes and IFN-inducible genes such as BATF2, CXCL10 and ANKRD22 between IFN+ and IFN- patients. This indicated that the regulation of the gene expression in IFN+ and IFN- TB patients does not happen through increased or decreased expression of IFN-α,-β or -γ signaling molecules.

Investigation into the biological differences between the IFN+ and IFN- TB patients

The GSEA-based stratification of the individuals into IFN- and IFN+ groups was shown to be biologically relevant using two independent datasets: one containing samples from individuals after and before influenza vaccination and one which contained individuals suffering from sepsis. In both studies I identified IFN+ and IFN- individuals. Moreover, I compared the IFN status of the IFN- and IFN+ volunteers after the FLUAD® vaccination with the measured total levels of IFN-inducible CCL2 and CXCL2 cytokines in their blood and showed that there is a significant difference in the levels of those cytokines between IFN+ and IFN- donors, which showed that the observations made on the transcriptomic level have their functional consequence in blood.

Comparison of the biosignatures of the IFN- and IFN+ TB patients

I identified the biosignatures of IFN+ and IFN- TB patients using ML methods. First, I created multiple RF models trained to detect the TB patients among healthy people, people with OD and all non-TB individuals. I derived various sizes of biosignatures of both IFN+ an IFN- TB patients and tested the performance of the biosignatures in identifying TB patients depending on their size (number

of included transcripts). The optimal size of IFN+ biosignature was 20 transcripts, while the optimal size of the IFN- biosignature was 50 transcripts. Next, using the training MDS I derived the gene signatures of IFN+ and IFN- TB patients. One of the features of an efficient biosignature of any disease is that it should detect all forms of this disease among healthy people as well as among people suffering of OD. The identified biosignatures of TB IFN+ and TB IFN- patients consisted of varying numbers of genes among which only one transcript was present in both biosignatures: ANKRD22. Among the transcripts present in the TB IFN+ signature 6 transcripts, and among the transcripts present in the TB IFN- signature 9 transcripts overlapped with the 53-transcript biosignature of TB identified by Kaforou et al. in 2013.

The derived signatures have been tested on the test MDS as well as on two independent datasets, one including healthy and TB patients from China (Cai et al., 2014) and one including healthy, TB and sarcoidosis patients recruited in London (Blankley, Graham, Levin, et al., 2016). The signature derived from IFN+ TB patients was highly sensitive and specific towards IFN+ TB patients, however its performance was significantly worse in identification of IFN- TB patients among HCs and in particular among patients with OD. On the other hand, the IFN- TB biosignature presented slightly lower AUC values for identification of TB patients overall, but the classification based on it was characterized by similar sensitivity and specificity of detection of IFN- and IFN+ patients among HCs, OD or all non-TB individuals. The exception was the detection of TB patients among sarcoidosis patients, in whom only IFN+ TB biosignature detected the TB patients correctly. The obtained results showed that (i) the IFN+ and IFN- TB patients are characterized by different biosignatures, and (ii) that the IFN- TB biosignature can be used to detect IFN+ TB patients with better outcome than the other way round, (iii) that even though the TB IFN- signature is more universal, it fails to differentiate between IFN+ TB patients and sarcoidosis patients.

As a control for the method of identification of IFN- and IFN+ patients as well as for the specificity of the derived biosignatures I identified a 20-transcript sepsis IFN+ signature and a 50-transcript sepsis IFN- biosignature. Interestingly, the TB patients could also be identified on the basis of sepsis biosignatures with moderate accuracy. However, the TB biosignatures were highly specific and did not correctly classify the sepsis patients. This indicated that the sepsis IFN+ signature is in part TB-specific, but also partially IFN-response specific. This might cause false negative results when applied to patients without a strong IFN response.

Analysis of the concordance of the gene expression between IFN+ and IFN- patients

The expression of several genes, e.g. CD274 and CD273, was strikingly different between IFN+ and IFN- TB patients which I have shown using disco.score. The CD274 and CD273 genes encode programmed-death ligands 1 (PD-L1) and 2 (PD-L2). PD-L1 and PD-L2 are immunomodulatory molecules that act largely through interaction with PD-1 receptor. PD1 interacts with PD-L1 and PD-

L2 delivering inhibitory signals to regulate T-cell and other responses (McNab et al., 2011). In TB it has been shown that antibodies blocking PD-L1, PD-L2 enhanced Mtb antigen-specific IFN-γ responses and CD8+ T cell cytotoxicity from peripheral blood and pleural fluid mononuclear cells (Hassan, Akram, King, Dockrell, & Cliff, 2015; McNab et al., 2011; Alvarez et al., 2010; Trinath et al., 2012). It has been previously observed that among TB patients who in general have increased levels of CD274 there are individuals with surprisingly low levels of expression of this gene (McNab et al., 2011). However, those patients were considered exceptional. In this thesis I show, that this phenotype is present among majority of the published TB transcriptomic studies. Thus, rather than being outliers, these patients form a subgroup presenting a different activation profile in response to TB. In the study by McNab from 2011 it has been stated that those exceptionally low level of CD274 expression were identified in a patient who at the same time did not present pathology in lungs in spite of being diagnosed with TB. This corresponds well with the findings of the work presented here. Using datasets containing X-Ray studies of lungs of the TB patients whose blood has been profiled on microarrays I present that IFN+ status corresponds with high level of pathology in lungs. However, I do not approach the question of whether the high level of IFN signaling promotes disease burden increase or rather the higher level of pathology induces stronger IFN response. The answer to this question cannot be drawn from the data analyzed here.

Other genes with significantly different but still concordant expression between IFN+ and IFN- patients include PAR2, C1QC, CARMIL2 and CAPN5. All of these genes have been previously reported to play an important role in TB or in immunity. The C1QC is a complement cascade gene and is present in some of the previously published TB biosignatures (Kaforou et al., 2013). PAR2 has been recently described as important for the inhibition of Mtb growth (Chávez-Galán, Ramon-Luing, Carranza, Garcia, & Sada-Ovalle, 2017). The differential expression of those genes between IFN+ and IFN- patients could therefore be involved in the observed relationship between the IFN levels and lung pathology in TB patients.

Immune response profiles of the TB patients

Focusing on the differences in the IFN response among TB patients is an approach to the main question of this thesis, which is whether different patterns of the immune response in TB exist. IFN response is only one of the elements of immune response against TB and is strongly related to T cells and NK cells which are responsible for IFN production. By creating the correlation matrix of the eigengenes of the modules enriched in TB patients I show that indeed also the other elements of immune responses differ between patient subgroups. The main division is related to innate and adaptive immunity – in most cases, increased transcript levels of the genes in the modules related to innate immunity correlated with decreased transcript levels of the genes in modules involved in adaptive immunity. In analogy, decreased transcript levels of the innate immunity genes were correlated with

increased transcript levels of the adaptive immunity genes. However, the patterns of immune response presented by groups of individuals were not as simple as that. The response of monocytes, platelets, modules related to erythropoesis and to activated CD4+ T cells also presented great variability and altogether the investigated modules formed six patterns of gene expression activation including activation and suppression of transcripts involved in particular modules.

I tested if differences related to the adaptive and innate immunity activation are explained by the phase of the disease which an individual is undergoing and which corresponds to the time p.i.. Since such hypothesis cannot be tested on human data because the infection time is not known, I used dataset from a study conducted on macaques. Analysis of datasets collected from 38 Mtb infected macaques among which 16 developed active TB and 22 remained latently infected (Gideon et al., 2016) contradicts this possible explanation of the mentioned differences. The variability in IFN response is observed in those animals independent of their disease status and severity and has different strength independent of time p.i.. Even though the general trend of the development of IFN response between 20 and 40 day p.i., which corresponds with the establishment of adaptive immunity, is visible, the fate of IFN response is variable in both active TB and LTBI macaques. Some of the LTBI animals retain strong IFN response also in the later days p.i. whereas some of the animals suffering of TB present very weak IFN response even throughout the time of peaking adaptive immunity phase. This suggests that strong IFN response does not depend on a particular stage of the disease development, but rather that its dynamics is highly host-related.

Limitations of this study

The presented study is an insight into individual variability among TB patients and it is characterized by many limitations among which some have been included already in its assumptions. First and foremost, human cohorts present a marked challenge to study because apart from genetic variability also the conditions of human life and circumstances of the infection and disease development are uncontrollable and untraceable. We do not know with what doses of Mtb the host got infected, neither in which moment of life or under what temporary state of host's organism the infection happened. Even the time points of TB diagnosis strongly differ: while some patients could have been identified during systematic screening for TB, others only come to the clinics while already having developed severe pathology. Moreover, another portion of variability is related to experiment planning and conducting, technical variation and various samples processing methods used in different studies. For this reason, in the presented work I proposed multi-level validation of my results. Primarily, I used k-fold cross-validation on the training MDS containing 80% of the acquired samples. The remaining 20% of the samples remained untouched and served for independent testing. Last, the obtained results were validated on two independent validation datasets and the whole pipeline on datasets from different diseases. I did also compare the transcriptomics-based results with the results of different types of

measurements: cytokine levels in blood and X-ray based detection of lung pathology. I compared my findings to previously noted observations in TB showing that in many of the published studies patients with weaker or unexpected gene expression profiles have been observed as well as that the expression of IFN-related genes was linked with the outcome of TB.

Useful methods and data collections presented in this study

I applied the proposed normalization and analysis methods in several external datasets not only from TB but also sepsis and influenza studies. The suggested analysis framework was robust and in the future can be used in other multi-cohort studies. A useful dataset collection has been selected out of the published TB datasets and can be accessed on the website: http://bioinfo.mpiib-berlin.mpg.de/TBprofiles/. Additionally, a new set of IFN type I, IFN type II and IFN type I and II inducible genes have been created.

Outlook

An important message from this study, strengthened by the inclusion of multiple human cohort datasets, platforms and studies, is that the attempts to define TB signature may involve a significant bias if they do not account for individual variability between hosts. Variable outcomes, pathology and even drastically different consequences of Mtb infection imply that it is not enough to assign the patients with TB into one of the general classes: "sick" or "healthy". The representations of this disease can on transcriptomic level be similar to inflammatory, immunosuppressive or chronic disease and I suggest that this should also be accounted for in TB diagnosis.

5.1.2. Comparison of the response to TB among different mouse strains

In this study, with the help of my colleagues from the Department of Immunology and Microarray Core Facility of MPIIB I identified elements of the immune response to TB which are conserved and which are divergent between man and two different mouse strains. Importantly, I demonstrated that a highly susceptible mouse strain 129S2 mimics human active TB more closely than the resistant C57BL/6 strain.

The development of a novel comparison method for heterogeneous datasets

The achievements of this project include development of a universal method for comparison of heterogeneous datasets - disco.score.

Intending to compare transcriptomic datasets from different mouse models of TB with the patients' data I tested the existing approaches of assessing similarities in gene expression in different species. Since none of the methods resulted in biologically meaningful conclusions I created an algorithm, disco.score, directed at identification of concordantly and discordantly regulated genes in a set of heterogeneous datasets. The motivation behind developing the disco.score was to create a

comparison algorithm which assumptions would encompass not only existence of similarities but also discordances in the responses to threat evolved by different organisms. Identification of these differences can be accompanied by the observation of the diverse phenotypes presented by the organisms from which the datasets were derived. This way, the identification of concordance and discordance is a source of information helping to explain the occurring different outcomes of particular disease. This information cannot be derived in a straightforward way from other common approaches like differential expression analysis which are not designed to be used directly with heterogeneous datasets, i.e. derived from different experiment types. Complemented by GSEA, the algorithm developed in this thesis indicates which elements of immune response are regulated in a similar or in opposite way between the compared datasets.

The proposed method was validated on datasets comparing responses to the same immune system stimulations in mouse and man or in different human populations and to compare the human immune responses to TB and sarcoidosis, diseases with nearly identical gene expression profile. Ultimately, the method was applied to compare transcriptional responses to TB in human and murine WB and monocytes. I acquired publicly available datasets from human WB and macrophages and my colleagues Lisa Scheuermann, Anca Dorhoi, Karin Hahnke (MPIIB, Department of Immunology) and Hans Mollenkopf (MPIIB, Microarray Core Facility) performed experiments to generate comparable datasets from WB of two mouse strains differing in susceptibility and frequently used to model TB as well as from murine macrophages and human macrophage-like THP-1 cells.

Characteristics of the disco.score

Disco.score can be applied to any number of gene expression datasets with calculated p-values and $\log_2$FC values for differential gene expression, but at a time it compares two datasets only. Instead of evaluating the overall similarities between the datasets, it identifies the most concordant and discordant genes or gene modules, which is reasonable in the light of evolutionary principles which imply that parts of the original systems, for example of the immune system, remain conserved and parts diverge over time.

Since disco.score can be calculated for every pair of heterologous genes present in the compared datasets it circumvents the bias of arbitrary gene choice. Lastly, the score is not restricted to be used in the comparison of immune system stimulation events, but it includes the possibility of using any gene sets of interest (e.g. GO terms or self-created gene modules).

Even though disco.score has been designed for the purpose of identifying similar and different elements of immune response between mouse and man, it can also be used in different types of analysis. For example, in the studies focusing on a particular gene or gene set, disco.score can serve to compare the expression regulation of the genes of interest among different conditions. The score can be used to

create novel gene sets – e.g. top ranking genes in disco.score can be combined in a new module characterizing a particular disease model.

Disco.score accuracy highly depends on the quality of the investigated data and availability of gene annotation. The algorithm does not allow correction for cell numbers which influence interpretation of the transcriptional studies and should be performed independently. Lack of an analytical distribution of disco.score means that no direct p-value can be derived for the score.

Comparison of the human datasets with two mouse models of TB using disco.score

Using disco.score I inspected the comparison of the gene expression patterns in two mouse strains, highly susceptible to TB 129S2 mouse and low susceptible C57BL/6 mouse upon the time course of infection. The number of concordances between them was increasing with time p.i. without any identified discordances. However, after the day 14 p.i. TB rapidly progressed in the susceptible mice. At the same time, discordant gene modules appeared between low susceptible vs highly susceptible mice and TB patients.

The highest number of similarities between the transcriptional profiles of patient cohorts from different geographical locations in Africa and mice was detected at day 21 p.i. The similarities encompassed IFN response, innate immunity mechanisms and B-cell signaling. At day 21 p.i. very few gene modules were discordant between the highly susceptible 129S2 mouse strain and TB patients. In contrast, significant number of modules remained discordant between the low susceptible C57BL/6 strain and TB patients. Those numbers reflected the fact, that TB patients as well as highly susceptible mice 21 days p.i. suffered from active TB, while the low susceptible C57BL/6 mice remained asymptomatic. This suggests that active TB is more accurately mimicked by the susceptible 129S2 mouse.

Investigation into the differences underlying the observed discordance in gene expression patterns of the C57BL/6 mouse strain and man

The discordances present at the 21st day p.i. between the human cohorts or highly susceptible 129S2 mice and low susceptible C57BL/6 mice were related to T cell functions. I inspected the regulation of the genes present in those modules in the course of infection and identified 16 of them as responsible for the discordance between 129S2 and human vs C57BL/6 gene expression. The 16 genes were regulated in opposite directions in the highly susceptible mouse vs low susceptible mouse and man at every time point p.i. T cell proliferation differences in lungs of susceptible mouse strain I/St and resistant mouse strain A/Sn have already been described, however without indicating genes responsible for this phenomenon (Eruslanov et al., 2004). Another study performed on susceptible and resistant macaque lineages also points out those differences (Javed et al., 2016). In the TB-susceptible macaques T-cell related genes were down-regulated at week 6 p.i. when animals had lost 10% body weight. The

genes CD28, CD3E and T-cell co-stimulatory molecules down-regulated in the macaques were at the same time identified as discordant by disco.score between susceptible and resistant mouse strains. It is therefore tempting to speculate that the 16 identified genes play a crucial role in acquired susceptibility and resistance in TB.

Interpretation of the obtained comparison results

The results of comparison of human and murine WB transcriptome need to be interpreted carefully, with attention paid to the fact that they might be influenced by the variation in the composition of human and murine blood. Apart from the differences in the healthy organisms, the cell counts change upon infection. For example, the CD4+ and CD8+ cell counts decrease in blood of TB patients compared to healthy individuals (Berry et al., 2010). 33 genes have been previously identified as disease-associated in a study on gene expression in sorted CD4+ and CD8+ T cell populations from TB patients, LTBI and controls (Jacobsen et al., 2011). This set of genes was enriched in JAK-STAT signaling pathway. I have found several of those genes present in the enriched module "Hallmark IL2 STAT5 signaling" identified as concordant in comparisons of South African and murine WB data from C57BL/6 and 129S2 strains.

Conclusion from comparing the high- and low susceptible mouse strains with human datasets

The conclusion of analysis of the two murine models´ similarity to human TB is that out of the investigated time points, mice at the time point of 21 days p.i. most resemble TB patients. At this time point the highly susceptible 129S2, but not the low susceptible C57BL/6 mice closely mimic genetic response in WB of the patients. However, since I did not investigate late time points p.i. of C57BL/6 mice, I cannot exclude that the WB transcriptional profile of these mice becomes more similar to human TB WB profile with time.

The obtained results demonstrate that depending on the animal model used, gene expression may variably mirror human disease and should be taken into account in translational studies, e.g. drug or vaccine tests. I also showed that the gene expression regulation in the macrophages upon Mtb infection is similar in mouse and man.

Outlook

This study comparing two profoundly different murine models of TB emphasizes the need to identify the best-fit animal model as correlate for a particular human disease. Disco.score as a straightforward, robust and simple method is the first step in this direction. It can help us to understand which elements of gene expression regulation are widely conserved between man and model organisms; which can be used in translational research and are species-specific and give low chances of translation to human.

For the field of TB research in which I gained deep insights during my doctoral studies I see a need of creating a compendium of discordant and concordant gene expression profiles between human cohorts and the animal models used to mimic TB. It would vastly help the scientists intending to conduct translational research to pick the right model for the specific topic investigated by them, for example T-cell response or innate immunity. The first step to create such compendium is to perform reproducible experiments and publish the obtained data. As soon as the datasets from broad range of murine and else animal models used to study TB is available in the public repositories, they can be compared with the human datasets for example with use of disco.score.

5.2. THE OUTLOOK OF THIS THESIS

TB remains a challenge for clinicians and researchers despite over 100 years of research since Mtb has been identified. Part of this challenge is related to the variability in the outcome of Mtb infection. In my doctoral work I looked at the variability in host responses to TB from different angles: investigating the differences between individual patients as well as searching for changes in the gene expression that explain different outcomes of Mtb infection in various animal models. I have shown that the inter-individual variability is strongly related to the extent of IFN signaling in human hosts. The variability between different murine models of TB was linked with the discordant expression of T-cell co-receptor genes.

These two results point out the mechanisms of high interest and high complexity which are extensively studied in the field of TB. My work brings additional insights into the background of phenotypic differences between highly and low susceptible murine models of TB and describes correlates of IFN signaling in TB patients. Altogether it is another tiny piece in the big picture of the complexity of TB. The described results are shown to be meaningful when seen in the context of other studies.

In my thesis I have developed and proposed methods which can contribute to further animal model studies of TB and other diseases as well as to analyze the individual variability of host responses to pathogen infection.

6. ACKNOWLEDGEMENTS

This work was supported by the German Federal Ministry of Education and Research (BMBF) funded consortium InfectControl 2020 (project Trans-sectoral research platform TFP-TV5, grant ID 03ZZ0802E) and ZIBI Graduate School Berlin.

My PhD was a very good and a highly forming time. I would like to also thank the people who personally helped me along it.

In the first place, I am thankful to Professor Stefan H. E. Kaufmann who accepted me as PhD student. He has been an inspiring example of a responsible boss, dedicated scientist, and bright advisor. I am thankful for his scientific advice and feel honored to have worked in his group.

Most of what I learned during my PhD, I learned thanks to and from my direct supervisor January Weiner. He has been a dedicated teacher, always happy to hear questions and doubts; a critical reviewer; a collaborator open to my ideas and a mine of knowledge. He gave me a set of tools and of principles which I will use in my scientific career. Dziękuję!

Barbara Broeker, Ulrik Stervbo, Stefan Kaufmann and January Weiner met me every year to discuss the progress of my projects. I thank them for their effort, sharing their knowledge, valuable hints and criticism.

I am very thankful to Lisa Scheuermann, Anca Dorhoi, Karin Hahnke and Hans Mollenkopf for fruitful discussions and collaboration. They performed the experiments described in this thesis and were fantastic co-authors.

I also thank my other collaborators and/or co-authors as well as the scientists who gave me important advice or comments regarding my work: Anouk Platteel, Raik Otto, Joanna Zyla, Haipong Liu, Macarena Beigier, Laura Lozza, Jeroen Maertzdorf, Gayle McEwen, Arturo Zychlinsky, Piotr Sobczyk, Michal Kusibab, Frida Arrey and Krzysztof Konina.

The Kaufmann Lab surrounded me with enthusiastic working environment, kindness and many stimulating discussions which I thank them for. I very much appreciate unfailing support of Souraya Sibaei, Katja Grunow and Robert Golinski.

The PhD coordinators Andreas Schmidt, Juliane Kofer and Susann Beetz enthusiastically supported my educational activities and scientific initiatives. I am very thankful to them for being always ready to help with any sorts of problems.

I thank Kevin Barnett and Garth Burn for reading and correcting English in this monograph.

I thank my reviewers: Stefan H.E. Kaufmann, Arturo Zychlinsky and Barbara Broeker for reviewing my work, as well as the other members of my Doctoral Committee: Edda Klipp and Benedikt Beckmann for participating in my doctoral examination.

Last, I would like to thank my wonderful family, boyfriend and friends for creating an environment in which I can grow. It makes me incredibly happy to share my initiatives, plans and daily life with you. It is also invaluable to be able to trust you with my struggles.

7. BIBLIOGRAPHY

Adams, J. U. (2014). *Essentials of Cell Biology*. (C. O'Connor, Ed.). Retrieved from http://onlinebooks.library.upenn.edu/webbin/book/lookupid?key=olbp65192

Alsina, L., Israelsson, E., Altman, M. C., Dang, K. K., Ghandil, P., Israel, L., … Chaussabel, D. (2014). A narrow repertoire of transcriptional modules responsive to pyogenic bacteria is impaired in patients carrying loss-of-function mutations in MYD88 or IRAK4. *Nature Immunology*, *15*(12), 1134–1142. https://doi.org/10.1038/ni.3028

Alvarez, I. B., Pasquinelli, V., Jurado, J. O., Abbate, E., Musella, R. M., de la Barrera, S. S., & García, V. E. (2010). Role Played by the Programmed Death-1–Programmed Death Ligand Pathway during Innate Immunity against *Mycobacterium tuberculosis*. *The Journal of Infectious Diseases*, *202*(4), 524–532. https://doi.org/10.1086/654932

Anderson, S. T., Kaforou, M., Brent, A. J., Wright, V. J., Banwell, C. M., Chagaluka, G., … Eley, B. (2014). Diagnosis of Childhood Tuberculosis and Host RNA Expression in Africa. *New England Journal of Medicine*, *370*(18), 1712–1723. https://doi.org/10.1056/NEJMoa1303657

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Athale, S., Banchereau, R., Thompson-Snipes, L., Wang, Y., Palucka, K., Pascual, V., & Banchereau, J. (2017). Influenza vaccines differentially regulate the interferon response in human dendritic cell subsets. *Science Translational Medicine*, *9*(382), eaaf9194. https://doi.org/10.1126/scitranslmed.aaf9194

Bach, E. A., Aguet, M., & Schreiber, R. D. (1997). The IFNγ Receptor: A Paradigm for Cytokine Receptor Signaling. *Annual Review of Immunology*, *15*(1), 563–591. https://doi.org/10.1146/annurev.immunol.15.1.563

Bachman, J. (2013). Reverse-Transcription PCR (RT-PCR). In *Methods in enzymology* (Vol. 530, pp. 67–74). https://doi.org/10.1016/B978-0-12-420037-1.00002-6

Banzhoff, A., Pellegrini, M., Del Giudice, G., Fragapane, E., Groth, N., & Podda, A. (2008). MF59-adjuvanted vaccines for seasonal and pandemic influenza prophylaxis. *Influenza and Other Respiratory Viruses*, *2*(6), 243–249. https://doi.org/10.1111/j.1750-2659.2008.00059.x

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., … Gifford, D. K. (2003). Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, *21*(11), 1337–1342. https://doi.org/10.1038/nbt890

Basu, R. K., Standage, S. W., Cvijanovich, N. Z., Allen, G. L., Thomas, N. J., Freishtat, R. J., … Wong, H. R. (2011). Identification of candidate serum biomarkers for severe septic shock-associated kidney injury via microarray. *Critical Care*, *15*(6), R273. https://doi.org/10.1186/cc10554

Bax, H. I., Freeman, A. F., Ding, L., Hsu, A. P., Marciano, B., Kristosturyan, E., … Sampaio, E. P. (2013). Interferon Alpha Treatment of Patients with Impaired Interferon Gamma Signaling. *Journal of Clinical Immunology*, *33*(5), 991–1001. https://doi.org/10.1007/s10875-013-9882-5

Beamer, G. L., & Turner, J. (2005). Murine models of susceptibility to tuberculosis. *Archivum Immunologiae et Therapiae Experimentalis*, *53*(6), 469–483.

Behar, S. M., & Boom, W. H. (2017). Unconventional T Cells. In *Handbook of Tuberculosis* (pp. 157–183). Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. https://doi.org/10.1002/9783527611614.ch24

Berry, M. P. R., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A. A., Oni, T., … O'Garra, A. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature*, *466*(7309), 973–977. https://doi.org/10.1038/nature09247

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, *69*(3), 89–95. https://doi.org/10.1067/mcp.2001.113989

Blankley, S., Graham, C. M., Levin, J., Turner, J., Berry, M. P. R., Bloom, C. I., … O'Garra, A. (2016). A 380-gene meta-signature of active tuberculosis compared with healthy controls. *European Respiratory Journal*, *47*(6), 1873–1876. https://doi.org/10.1183/13993003.02121-2015

Blankley, S., Graham, C. M., Turner, J., Berry, M. P. R., Bloom, C. I., Xu, Z., … O'Garra, A. (2016). The Transcriptional Signature of Active Tuberculosis Reflects Symptom Status in Extra-Pulmonary and Pulmonary Tuberculosis. *PLOS ONE*, *11*(10), e0162220. https://doi.org/10.1371/journal.pone.0162220

Blomgran, R., & Ernst, J. D. (2011). Lung Neutrophils Facilitate Activation of Naive Antigen-Specific CD4+ T Cells during Mycobacterium tuberculosis Infection. *The Journal of Immunology*, *186*(12), 7110–7119. https://doi.org/10.4049/jimmunol.1100001

Bloom, C. I., Graham, C. M., Berry, M. P. R., Rozakeas, F., Redford, P. S., Wang, Y., … O'Garra, A. (2013). Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. *PloS One*, *8*(8), e70630. https://doi.org/10.1371/journal.pone.0070630

Bloom, C. I., Graham, C. M., Berry, M. P. R., Wilkinson, K. A., Oni, T., Rozakeas, F., … O'Garra, A. (2012). Detectable Changes in The Blood Transcriptome Are Present after Two Weeks of Antituberculosis Therapy. *PLoS ONE*, *7*(10), e46191. https://doi.org/10.1371/journal.pone.0046191

Bold, T. D., Banaei, N., Wolf, A. J., & Ernst, J. D. (2011). Suboptimal Activation of Antigen-Specific CD4+ Effector Cells Enables Persistence of M. tuberculosis In Vivo. *PLoS Pathogens*, *7*(5), e1002063. https://doi.org/10.1371/journal.ppat.1002063

Boxx, G. M., & Cheng, G. (2016). The Roles of Type I Interferon in Bacterial Infection. *Cell Host & Microbe*, *19*(6), 760–769. https://doi.org/10.1016/j.chom.2016.05.016

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cagliani, R., & Sironi, M. (2013). Pathogen-Driven Selection in the Human Genome. *International Journal of Evolutionary Biology*, *2013*, 1–6. https://doi.org/10.1155/2013/204240

Cai, Y., Yang, Q., Tang, Y., Zhang, M., Liu, H., Zhang, G., … Chen, X. (2014). Increased Complement C1q Level Marks Active Disease in Human Tuberculosis. *PLoS ONE*, *9*(3), e92340. https://doi.org/10.1371/journal.pone.0092340

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., … Large Scale Collab. Res. Program, I. and H. R. to I. (2005). A network-based analysis of systemic inflammation in humans. *Nature*, *437*(7061), 1032–1037. https://doi.org/10.1038/nature03985

Carow, B., Ye, X. qun, Gavier-Widén, D., Bhuju, S., Oehlmann, W., Singh, M., … Rottenberg, M. E. (2011). Silencing suppressor of cytokine signaling-1 (SOCS1) in macrophages improves Mycobacterium tuberculosis control in an interferon-gamma (IFN-gamma)-dependent manner. *The Journal of Biological Chemistry*, *286*(30), 26873–26887. https://doi.org/10.1074/jbc.M111.238287

Caruso, A. M., Serbina, N., Klein, E., Triebold, K., Bloom, B. R., & Flynn, J. L. (1999). Mice deficient in CD4 T cells have only transiently diminished levels of IFN-gamma, yet succumb to tuberculosis. *Journal of Immunology (Baltimore, Md. : 1950)*, *162*(9), 5407–5416. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10228018

Chackerian, A., Alt, J., Perera, V., & Behar, S. M. (2002). Activation of NKT cells protects mice from tuberculosis. *Infection and Immunity*, *70*(11), 6302–6309. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12379709

Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., … Pascual, V. (2008). A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity*, *29*(1), 150–164. https://doi.org/10.1016/j.immuni.2008.05.012

Chávez-Galán, L., Ramon-Luing, L., Carranza, C., Garcia, I., & Sada-Ovalle, I. (2017). Lipoarabinomannan Decreases Galectin-9 Expression and Tumor Necrosis Factor Pathway in Macrophages Favoring Mycobacterium tuberculosis Intracellular Growth. *Frontiers in Immunology*, *8*, 1659. https://doi.org/10.3389/fimmu.2017.01659

Churchyard, G., Kim, P., Shah, N. S., Rustomjee, R., Gandhi, N., Mathema, B., … Cardenas, V. (2017). What We Know About Tuberculosis Transmission: An Overview. *The Journal of Infectious Diseases*, *216*(suppl_6), S629–S635. https://doi.org/10.1093/infdis/jix362

Cliff, J. M., Lee, J.-S., Constantinou, N., Cho, J.-E., Clark, T. G., Ronacher, K., … Dockrell, H. M. (2013). Distinct Phases of Blood Gene Expression Pattern Through Tuberculosis Treatment Reflect Modulation of the Humoral Immune Response. *The Journal of Infectious Diseases*, *207*(1), 18–29. https://doi.org/10.1093/infdis/jis499

Cooper, A. M. (2009). Cell-Mediated Immune Responses in Tuberculosis. *Annual Review of Immunology*, *27*(1), 393–422. https://doi.org/10.1146/annurev.immunol.021908.132703

Cooper, A. M. (2014). Mouse model of tuberculosis. *Cold Spring Harbor Perspectives in Medicine*, *5*(2), a018556. https://doi.org/10.1101/cshperspect.a018556

Cooper, A. M., Adams, L. B., Dalton, D. K., Appelberg, R., & Ehlers, S. (2002). IFN-γ and NO in mycobacterial disease: new jobs for old hands. *Trends in Microbiology*, *10*(5), 221–226. https://doi.org/10.1016/S0966-842X(02)02344-2

Cooper, A. M., Dalton, D. K., Stewart, T. A., Griffin, J. P., Russell, D. G., & Orme, I. M. (1993). Disseminated tuberculosis in interferon gamma gene-disrupted mice. *The Journal of Experimental Medicine*, *178*(6), 2243–2247. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8245795

Cooper, A. M., Magram, J., Ferrante, J., & Orme, I. M. (1997). Interleukin 12 (IL-12) is crucial to the development of protective immunity in mice intravenously infected with mycobacterium tuberculosis. *The Journal of Experimental Medicine*, *186*(1), 39–45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9206995

Davenport, E. E., Burnham, K. L., Radhakrishnan, J., Humburg, P., Hutton, P., Mills, T. C., … Knight, J. C. (2016). Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. *The Lancet. Respiratory Medicine*, *4*(4), 259–271. https://doi.org/10.1016/S2213-2600(16)00046-1

Davidson, S., Crotta, S., McCabe, T. M., Wack, A., Peiris, J. S., Cheung, C. Y., … Brody, S. L. (2014). Pathogenic potential of interferon αβ in acute influenza infection. *Nature Communications*, *5*, 574–584. https://doi.org/10.1038/ncomms4864

Davis, S., & Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England)*, *23*(14), 1846–1847. https://doi.org/10.1093/bioinformatics/btm254

Dawany, N., Showe, L. C., Kossenkov, A. V., Chang, C., Ive, P., Conradie, F., … Montaner, L. J. (2014). Identification of a 251 Gene Expression Signature That Can Accurately Detect M. tuberculosis in Patients with and without HIV Co-Infection. *PLoS ONE*, *9*(2), e89925. https://doi.org/10.1371/journal.pone.0089925

Desvignes, L., & Ernst, J. D. (2009). Interferon-γ-Responsive Nonhematopoietic Cells Regulate the Immune Response to Mycobacterium tuberculosis. *Immunity*, *31*(6), 974–985. https://doi.org/10.1016/j.immuni.2009.10.007

Desvignes, L., Wolf, A. J., & Ernst, J. D. (2012). Dynamic Roles of Type I and Type II IFNs in Early Infection with Mycobacterium tuberculosis. *The Journal of Immunology*, *188*(12), 6205–6215. https://doi.org/10.4049/jimmunol.1200255

Dobbs, T., & Kimmerling, M. (2008). Mycobacterium tuberculosis. In AIDS Therapy E-book. In *AIDS Therapy E-book.* Philadelphia: Churchill Livingstone/Elsevier.

Domaszewska, T., Scheuermann, L., Hahnke, K., Mollenkopf, H., Dorhoi, A., Kaufmann, S. H. E., & Weiner, J. (2017). Concordant and discordant gene expression patterns in mouse strains identify best-fit animal model for human tuberculosis. *Scientific Reports*, *7*(1), 12094. https://doi.org/10.1038/s41598-017-11812-x

Donovan, M. L., Schultz, T. E., Duke, T. J., & Blumenthal, A. (2017). Type I Interferons in the Pathogenesis of Tuberculosis: Molecular Drivers and Immunological Consequences. *Frontiers in Immunology*, *8*, 1633. https://doi.org/10.3389/fimmu.2017.01633

Dorhoi, A., Iannaccone, M., Farinacci, M., Faé, K. C., Schreiber, J., Moura-Alves, P., … Yakhini, Z. (2013). MicroRNA-223 controls susceptibility to tuberculosis by regulating lung neutrophil recruitment. *Journal of Clinical Investigation*, *123*(11), 4836–4848. https://doi.org/10.1172/JCI67604

Dorhoi, A., & Kaufmann, S. H. E. (2015). Versatile myeloid cell subsets contribute to tuberculosis-associated inflammation. *European Journal of Immunology*, *45*(8), 2191–2202. https://doi.org/10.1002/eji.201545493

Dorhoi, A., Yeremeev, V., Nouailles, G., Weiner, J., Jörg, S., Heinemann, E., … Kaufmann, S. H. E. (2014). Type I IFN signaling triggers immunopathology in tuberculosis-susceptible mice by modulating lung phagocyte dynamics. *European Journal of Immunology*, *44*(8), 2380–2393. https://doi.org/10.1002/eji.201344219

Downing, G. J. (2000). *Biomarkers and surrogate endpoints : clinical research and applications : proceedings of the NIH-FDA conference held on 15-16 April 1999 in Bethesda, Maryland, USA*. (N. I. of H. (U.S.) & United States. Food and Drug Administration., Eds.). Amsterdam: Elsevier. Retrieved from http://www.worldcat.org/title/biomarkers-and-surrogate-endpoints-clinical-research-and-applications-proceedings-of-the-nih-fda-conference-held-on-15-16-april-1999-in-bethesda-maryland-usa/oclc/247805078

Driver, E. R., Ryan, G. J., Hoff, D. R., Irwin, S. M., Basaraba, R. J., Kramnik, I., & Lenaerts, A. J. (2012). Evaluation of a Mouse Model of Necrotic Granuloma Formation Using C3HeB/FeJ Mice for Testing of Drugs against Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy*, *56*(6), 3181–3195. https://doi.org/10.1128/AAC.00217-12

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., & Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics*, *21*(1s), 10–14. https://doi.org/10.1038/4434

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics (Oxford, England)*, *21*(16), 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11752295

Egen, J. G., Rothfuchs, A. G., Feng, C. G., Horwitz, M. A., Sher, A., & Germain, R. N. (2011). Intravital Imaging Reveals Limited Antigen Presentation and T Cell Effector Function in Mycobacterial Granulomas. *Immunity*, *34*(5), 807–819. https://doi.org/10.1016/j.immuni.2011.03.022

Egen, J. G., Rothfuchs, A. G., Feng, C. G., Winter, N., Sher, A., & Germain, R. N. (2008). Macrophage and T Cell Dynamics during the Development and Disintegration of Mycobacterial Granulomas. *Immunity*, *28*(2), 271–284. https://doi.org/10.1016/j.immuni.2007.12.010

Elemam, N. M., Hannawi, S., & Maghazachi, A. A. (2017). Innate Lymphoid Cells (ILCs) as Mediators of Inflammation, Release of Cytokines and Lytic Molecules. *Toxins*, *9*(12). https://doi.org/10.3390/toxins9120398

Eruslanov, E. B., Lyadova, I. V, Kondratieva, T. K., Majorov, K. B., Scheglov, I. V, Orlova, M. O., & Apt, A. S. (2005). Neutrophil responses to Mycobacterium tuberculosis infection in genetically susceptible and resistant mice. *Infection and Immunity*, *73*(3), 1744–1753. https://doi.org/10.1128/IAI.73.3.1744-1753.2005

Eruslanov, E. B., Majorov, K. B., Orlova, M. O., Mischenko, V. V, Kondratieva, T. K., Apt, A. S., & Lyadova, I. V. (2004). Lung cell responses to M. tuberculosis in genetically susceptible and resistant mice following intratracheal challenge. *Clinical and Experimental Immunology*, *135*(1), 19–28. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14678260

Feng, C. G., Kaviratne, M., Rothfuchs, A. G., Cheever, A., Hieny, S., Young, H. A., … Sher, A. (2006). NK cell-derived IFN-gamma differentially regulates innate resistance and neutrophil response in T cell-deficient hosts infected with Mycobacterium tuberculosis. *Journal of Immunology (Baltimore, Md. : 1950)*, *177*(10), 7086–7093. https://doi.org/10.4049/JIMMUNOL.177.10.7086

FIND. Because diagnostics matters. (2018). Retrieved May 29, 2018, from https://www.finddx.org/target-product-profiles/#tb

Fiusa, M. M. L., Carvalho, B. S., Hubert, R. M. E., Souza, W., Lopes-Cendes, I., Annichino-Bizzacchi, J. M., & De Paula, E. V. (2014). A Meta-Analysis of Gene Expression Studies in Severe Sepsis and Septic Shock. *Blood*, *124*(21). Retrieved from http://www.bloodjournal.org/content/124/21/2741?sso-checked=true

Flynn, J. L., & Chan, J. (2001). Immunology of tuberculosis. *Annual Review of Immunology*, *19*(1), 93–

129. https://doi.org/10.1146/annurev.immunol.19.1.93
Flynn, J. L., Chan, J., Triebold, K. J., Dalton, D. K., Stewart, T. A., & Bloom, B. R. (1993). An essential role for interferon gamma in resistance to Mycobacterium tuberculosis infection. *The Journal of Experimental Medicine*, *178*(6), 2249–2254. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7504064
Foell, D., Wittkowski, H., Kessel, C., Lüken, A., Weinhage, T., Varga, G., … Roth, J. (2013). Proinflammatory S100A12 Can Activate Human Monocytes via Toll-like Receptor 4. *American Journal of Respiratory and Critical Care Medicine*, *187*(12), 1324–1334. https://doi.org/10.1164/rccm.201209-1602OC
Fox, G. J., Barry, S. E., Britton, W. J., & Marks, G. B. (2013). Contact investigation for tuberculosis: a systematic review and meta-analysis. *European Respiratory Journal*, *41*(1), 140–156. https://doi.org/10.1183/09031936.00070812
Froeschle, J. E., Ruben, F. L., & Bloh, A. M. (2002). Immediate Hypersensitivity Reactions after Use of Tuberculin Skin Testing. *Clinical Infectious Diseases*, *34*(1), e12–e13. https://doi.org/10.1086/324587
Gallegos, A. M., van Heijst, J. W. J., Samstein, M., Su, X., Pamer, E. G., & Glickman, M. S. (2011). A Gamma Interferon Independent Mechanism of CD4 T Cell Mediated Control of M. tuberculosis Infection in vivo. *PLoS Pathogens*, *7*(5), e1002052. https://doi.org/10.1371/journal.ppat.1002052
Gideon, H. P., Skinner, J. A., Baldwin, N., Flynn, J. L., & Lin, P. L. (2016). Early Whole Blood Transcriptional Signatures Are Associated with Severity of Lung Inflammation in Cynomolgus Macaques with *Mycobacterium tuberculosis* Infection. *The Journal of Immunology*, *197*(12), 4817–4828. https://doi.org/10.4049/jimmunol.1601138
Giosue, S., Casarini, M., Alemano, L., Gallucio, G., Mattia, P., Pedicelli, G., … Ameglio, F. (1998). Effects of Aerosolized Interferon- α in Patients with Pulmonary Tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, *158*(4), 1156–1162. https://doi.org/10.1164/ajrccm.158.4.9803065
Godec, J., Tan, Y., Liberzon, A., Tamayo, P., Bhattacharya, S., Butte, A. J., … Haining, W. N. (2016). Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. *Immunity*, *44*(1), 194–206. https://doi.org/10.1016/j.immuni.2015.12.006
Golden, M. P., & Vikram, H. R. (2005). Extrapulmonary tuberculosis: an overview. *American Family Physician*, *72*(9), 1761–1768. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16300038
Govoni, G., Vidal, S., Gauthier, S., Skamene, E., Malo, D., & Gros, P. (1996). The Bcg/Ity/Lsh locus: genetic transfer of resistance to infections in C57BL/6J mice transgenic for the Nramp1 Gly169 allele. *Infection and Immunity*, *64*(8), 2923–2929. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8757814
Gupta, R. K., Lawn, S. D., Bekker, L.-G., Caldwell, J., Kaplan, R., & Wood, R. (2013). Impact of human immunodeficiency virus and CD4 count on tuberculosis diagnosis: analysis of city-wide data from Cape Town, South Africa. *The International Journal of Tuberculosis and Lung Disease*, *17*(8), 1014–1022. https://doi.org/10.5588/ijtld.13.0032
Gupta, U. D., & Katoch, V. M. (2005). Animal models of tuberculosis. *Tuberculosis*, *85*(5–6), 277–293. https://doi.org/10.1016/j.tube.2005.08.008
Hamilton, L. D. (2014). Introduction to Principal Component Analysis (PCA). Retrieved June 13, 2018, from http://www.lauradhamilton.com/introduction-to-principal-component-analysis-pca
Harari, A., Rozot, V., Enders, F. B., Perreau, M., Stalder, J. M., Nicod, L. P., … Pantaleo, G. (2011). Dominant TNF-α+ Mycobacterium tuberculosis–specific CD4+ T cell responses discriminate between latent infection and active disease. *Nature Medicine*, *17*(3), 372–376. https://doi.org/10.1038/nm.2299
Hassan, S. S., Akram, M., King, E. C., Dockrell, H. M., & Cliff, J. M. (2015). PD-1, PD-L1 and PD-L2 Gene Expression on T-Cells and Natural Killer Cells Declines in Conjunction with a Reduction in PD-1 Protein during the Intensive Phase of Tuberculosis Treatment. *PLOS ONE*, *10*(9), e0137646. https://doi.org/10.1371/journal.pone.0137646
Hatherill, M. (2011). Prospects for elimination of childhood tuberculosis: the role of new vaccines. *Archives of Disease in Childhood*, *96*(9), 851–856. https://doi.org/10.1136/adc.2011.214494

He, S. L., & Green, R. (2013). Northern blotting. *Methods in Enzymology*, *530*, 75–87. https://doi.org/10.1016/B978-0-12-420037-1.00003-8

Hopewell, P. C. (2017). Clinical Features of Tuberculosis. In *Handbook of Tuberculosis* (pp. 89–113). Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. https://doi.org/10.1002/9783527611614.ch37

Howes, A., Taubert, C., Blankley, S., Spink, N., Wu, X., Graham, C. M., … O'Garra, A. (2016). Differential Production of Type I IFN Determines the Reciprocal Levels of IL-10 and Proinflammatory Cytokines Produced by C57BL/6 and BALB/c Macrophages. *The Journal of Immunology*, *197*(7), 2838–2853. https://doi.org/10.4049/jimmunol.1501923

Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Jacobsen, M., Repsilber, D., Gutschmidt, A., Neher, A., Feldmann, K., Mollenkopf, H. J., … Kaufmann, S. H. E. (2007). Candidate biomarkers for discrimination between infection and disease caused by Mycobacterium tuberculosis. *Journal of Molecular Medicine*, *85*(6), 613–621. https://doi.org/10.1007/s00109-007-0157-6

Jacobsen, M., Repsilber, D., Kleinsteuber, K., Gutschmidt, A., Schommer-Leitner, S., Black, G., … Kaufmann, S. H. E. (2011). Suppressor of cytokine signaling-3 is affected in T-cells from tuberculosisTB patients. *Clinical Microbiology and Infection*, *17*(9), 1323–1331. https://doi.org/10.1111/j.1469-0691.2010.03326.x

Jagga, Z., & Gupta, D. (2015). Machine learning for biomarker identification in cancer research – developments toward its clinical application. *Personalized Medicine*, *12*(4), 371–387. https://doi.org/10.2217/pme.15.5

Javed, S., Marsay, L., Wareham, A., Lewandowski, K. S., Williams, A., Dennis, M. J., … Ostrand-Rosenberg, S. (2016). Temporal Expression of Peripheral Blood Leukocyte Biomarkers in a Macaca fascicularis Infection Model of Tuberculosis; Comparison with Human Datasets and Analysis with Parametric/Non-parametric Tools for Improved Diagnostic Biomarker Identification. *PLOS ONE*, *11*(5), e0154320. https://doi.org/10.1371/journal.pone.0154320

Kaforou, M., Wright, V. J., Oni, T., French, N., Anderson, S. T., Bangani, N., … Levin, M. (2013). Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study. *PLoS Medicine*, *10*(10), e1001538. https://doi.org/10.1371/journal.pmed.1001538

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10592173

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. https://doi.org/10.1093/nar/gkv1070

Kashino, S. S., Pollock, N., Napolitano, D. R., Rodrigues Jr, V., & Campos-Neto, A. (2008). Identification and characterization of *Mycobacterium tuberculosis* antigens in urine of patients with active pulmonary tuberculosis: an innovative and alternative approach of antigen discovery of useful microbial molecules. *Clinical & Experimental Immunology*, *153*(1), 56–62. https://doi.org/10.1111/j.1365-2249.2008.03672.x

Kaufmann, S. H. E. (2010). Future Vaccination Strategies against Tuberculosis: Thinking outside the Box. *Immunity*, *33*(4), 567–577. https://doi.org/10.1016/J.IMMUNI.2010.09.015

Kaufmann, S. H. E., Hussey, G., & Lambert, P.-H. (2010). New vaccines for tuberculosis. *Lancet (London, England)*, *375*(9731), 2110–2119. https://doi.org/10.1016/S0140-6736(10)60393-5

Kayagaki, N., Warming, S., Lamkanfi, M., Walle, L. Vande, Louie, S., Dong, J., … Dixit, V. M. (2011). Non-canonical inflammasome activation targets caspase-11. *Nature*, *479*(7371), 117–121. https://doi.org/10.1038/nature10558

Keller, C., Hoffmann, R., Lang, R., Brandau, S., Hermann, C., & Ehlers, S. (2006). Genetically Determined Susceptibility to Tuberculosis in Mice Causally Involves Accelerated and Enhanced Recruitment of Granulocytes. *Infection and Immunity*, *74*(7), 4295–4309.

https://doi.org/10.1128/IAI.00057-06
Kondratieva, T. K., Rubakova, E. I., Linge, I. A., Evstifeev, V. V., Majorov, K. B., & Apt, A. S. (2010). B Cells Delay Neutrophil Migration toward the Site of Stimulus: Tardiness Critical for Effective Bacillus Calmette-Guerin Vaccination against Tuberculosis Infection in Mice. *The Journal of Immunology*, *184*(3), 1227–1234. https://doi.org/10.4049/jimmunol.0902011
Kozakiewicz, L., Chen, Y., Xu, J., Wang, Y., Dunussi-Joannopoulos, K., Ou, Q., … Chan, J. (2013). B Cells Regulate Neutrophilia during Mycobacterium tuberculosis Infection and BCG Vaccination by Modulating the Interleukin-17 Response. *PLoS Pathogens*, *9*(7), e1003472. https://doi.org/10.1371/journal.ppat.1003472
Kramnik, I., Demant, P., & Bloom, B. B. (1998). Susceptibility to tuberculosis as a complex genetic trait: analysis using recombinant congenic strains of mice. *Novartis Foundation Symposium*, *217*, 120-31; discussion 132-7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9949805
Kramnik, I., Dietrich, W. F., Demant, P., & Bloom, B. R. (2000). Genetic control of resistance to experimental infection with virulent Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, *97*(15), 8560–8565. https://doi.org/10.1073/pnas.150227197
Krikorian, G., Marshall, W. H., Simmons, S., & Stratton, F. (1975). Counts and characteristics of macrophage precursors in human peripheral blood. *Cellular Immunology*, *19*(1), 22–31. https://doi.org/10.1016/0008-8749(75)90288-9
Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05
Kühn, R., & Torres, R. M. (2002). Cre/ loxP Recombination System and Gene Targeting. In *Transgenesis Techniques* (Vol. 180, pp. 175–204). New Jersey: Humana Press. https://doi.org/10.1385/1-59259-178-7:175
Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559
Lawn, S. D. (2015). Advances in Diagnostic Assays for Tuberculosis. *Cold Spring Harbor Perspectives in Medicine*, *5*(12), a017806. https://doi.org/10.1101/cshperspect.a017806
Leek, J., Johnson, W., Parker, H., Fertig, E., Jaffe, A., Storey, J., … Torres, L. (2018). Surrogate Variable Analysis. R package version 3.28.0. Retrieved from https://bioconductor.org/packages/release/bioc/html/sva.html
Lesho, E., Forestiero, F. J., Hirata, M. H., Hirata, R. D., Cecon, L., Melo, F. F., … Ooi, G. T. (2011). Transcriptional responses of host peripheral blood cells to tuberculosis infection. *Tuberculosis*, *91*(5), 390–399. https://doi.org/10.1016/j.tube.2011.07.002
Leung, M. K. K., Delong, A., Alipanahi, B., & Frey, B. J. (2016). Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. *Proceedings of the IEEE*, *104*(1), 176–197. https://doi.org/10.1109/JPROC.2015.2494198
Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., … Pulendran, B. (2014). Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature Immunology*, *15*(2), 195–204. https://doi.org/10.1038/ni.2789
Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3).
Liew, C.-C., Ma, J., Tang, H.-C., Zheng, R., & Dempsey, A. A. (2006). The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. *Journal of Laboratory and Clinical Medicine*, *147*(3), 126–132. https://doi.org/10.1016/J.LAB.2005.10.005
Lill, M., Kõks, S., Soomets, U., Schalkwyk, L. C., Fernandes, C., Lutsar, I., & Taba, P. (2013). Peripheral blood RNA gene expression profiling in patients with bacterial meningitis. *Frontiers in Neuroscience*, *7*, 33. https://doi.org/10.3389/fnins.2013.00033
Lin, S., Lin, Y., Nery, J. R., Urich, M. A., Breschi, A., Davis, C. A., … Snyder, M. P. (2014). Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences*, *111*(48), 201413624. https://doi.org/10.1073/pnas.1413624111
Liu, X., Jessen, W. J., Sivaganesan, S., Aronow, B. J., & Medvedovic, M. (2007). Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and ChIP-chip data. *BMC Bioinformatics*, *8*(1), 283. https://doi.org/10.1186/1471-2105-8-283
Lowe, D. M., Redford, P. S., Wilkinson, R. J., O'Garra, A., & Martineau, A. R. (2012). Neutrophils in

tuberculosis: friend or foe? *Trends in Immunology*, *33*(1), 14–25. https://doi.org/10.1016/j.it.2011.10.003

Lu, C., Wu, J., Wang, H., Wang, S., Diao, N., Wang, F., … Zhang, W. (2011). Novel biomarkers distinguishing active tuberculosis from latent infection identified by gene expression profile of peripheral blood mononuclear cells. *PloS One*, *6*(8), e24290. https://doi.org/10.1371/journal.pone.0024290

Macgregor, P. F., & Squire, J. A. (2002). Application of microarrays to the analysis of gene expression in cancer. *Clinical Chemistry*, *48*(8), 1170–1177. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12142369

Maertzdorf, J., Kaufmann, S. H. E., & Weiner, J. (2014). Toward a Unified Biosignature for Tuberculosis. In S. H. E. Kaufmann, E. Rubin, & A. Zumla (Eds.), *Tuberculosis* (pp. 183–196). New York: Cold Spring Harbor Laboratory Press.

Maertzdorf, J., Ota, M., Repsilber, D., Mollenkopf, H. J., Weiner, J., Hill, P. C., & Kaufmann, S. H. E. (2011). Functional Correlations of Pathogenesis-Driven Gene Expression Signatures in Tuberculosis. *PLoS ONE*, *6*(10), e26938. https://doi.org/10.1371/journal.pone.0026938

Maertzdorf, J., Repsilber, D., Parida, S. K., Stanley, K., Roberts, T., Black, G., … Kaufmann, S. H. E. (2011). Human gene expression profiles of susceptibility and resistance in tuberculosis. *Genes & Immunity*, *12*(1), 15–22. https://doi.org/10.1038/gene.2010.51

Maertzdorf, J., Weiner, J., Mollenkopf, H.-J., Bauer, T., Prasse, A., Müller-Quernheim, J., & Kaufmann, S. H. E. (2012). Common patterns and disease-related signatures in tuberculosis and sarcoidosis. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(20), 7853–7858. https://doi.org/10.1073/pnas.1121072109

Maglione, P. J., & Chan, J. (2009). How B cells shape the immune response against *Mycobacterium tuberculosis*. *European Journal of Immunology*, *39*(3), 676–686. https://doi.org/10.1002/eji.200839148

Maglione, P. J., Xu, J., & Chan, J. (2007). B cells moderate inflammatory progression and enhance bacterial containment upon pulmonary challenge with Mycobacterium tuberculosis. *Journal of Immunology (Baltimore, Md. : 1950)*, *178*(11), 7222–7234. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17513771

Manca, C., Tsenova, L., Freeman, S., Barczak, A. K., Tovey, M., Murray, P. J., … Kaplan, G. (2005). Hypervirulent *M. tuberculosis* W/Beijing Strains Upregulate Type I IFNs and Increase Expression of Negative Regulators of the Jak-Stat Pathway. *Journal of Interferon & Cytokine Research*, *25*(11), 694–701. https://doi.org/10.1089/jir.2005.25.694

Mansoori, D., Tavana, S., Mirsaeidi, M., Yazdanpanah, M., & Sohrabpour, H. (2002). The Efficacy of Interferon-α in the Treatment of Multidrug Resistant Tuberculosis. *Tanaffos*, *1*(3), 29–34.

Mayer-Barber, K. D., Andrade, B. B., Oland, S. D., Amaral, E. P., Barber, D. L., Gonzales, J., … Sher, A. (2014). Host-directed therapy of tuberculosis based on interleukin-1 and type I interferon crosstalk. *Nature*, *511*(7507), 99–103. https://doi.org/10.1038/nature13489

McNab, F. W., Berry, M. P. R., Graham, C. M., Bloch, S. A. A., Oni, T., Wilkinson, K. A., … O'Garra, A. (2011). Programmed death ligand 1 is over-expressed by neutrophils in the blood of patients with active tuberculosis. *European Journal of Immunology*, *41*(7), 1941–1947. https://doi.org/10.1002/eji.201141421

McNab, F. W., Ewbank, J., Rajsbaum, R., Stavropoulos, E., Martirosyan, A., Redford, P. S., … O'Garra, A. (2013). TPL-2-ERK1/2 signaling promotes host resistance against intracellular bacterial infection by negative regulation of type I IFN production. *Journal of Immunology (Baltimore, Md. : 1950)*, *191*(4), 1732–1743. https://doi.org/10.4049/jimmunol.1300146

Medina, & North. (2001). Resistance ranking of some common inbred mouse strains to Mycobacterium tuberculosis and relationship to major histocompatibility complex haplotype and Nramp1 genotype. *Immunology*, *93*(2), 270–274. https://doi.org/10.1046/j.1365-2567.1998.00419.x

Mestas, J., & Hughes, C. C. W. (2004). Of Mice and Not Men: Differences between Mouse and Human Immunology. *The Journal of Immunology*, *172*(5), 2731–2738. https://doi.org/10.4049/jimmunol.172.5.2731

Minion, J., Leung, E., Talbot, E., Dheda, K., Pai, M., & Menzies, D. (2011). Diagnosing tuberculosis with urine lipoarabinomannan: systematic review and meta-analysis. *European Respiratory*

Journal, *38*(6), 1398–1405. https://doi.org/10.1183/09031936.00025711
Mistry, R., Cliff, J. M., Clayton, C. L., Beyers, N., Mohamed, Y. S., Wilson, P. A., … Lukey, P. T. (2007). Gene-Expression Patterns in Whole Blood Identify Subjects at Risk for Recurrent Tuberculosis. *The Journal of Infectious Diseases*, *195*(3), 357–365. https://doi.org/10.1086/510397
Miyahira, A. (2012). Types of immune cells present in human PBMC. Retrieved June 13, 2018, from https://technical.sanguinebio.com/types-of-immune-cells-present-in-human-pbmc/
Monteiro, R. C., & van de Winkel, J. G. J. (2003). Ig A Fc Receptors. *Annual Review of Immunology*, *21*(1), 177–204. https://doi.org/10.1146/annurev.immunol.21.120601.141011
Moreira-Teixeira, L., Sousa, J., McNab, F. W., Torrado, E., Cardoso, F., Machado, H., … Saraiva, M. (2016). Type I IFN Inhibits Alternative Macrophage Activation during Mycobacterium tuberculosis Infection and Leads to Enhanced Protection in the Absence of IFN-γ Signaling. *Journal of Immunology (Baltimore, Md. : 1950)*, *197*(12), 4714–4726. https://doi.org/10.4049/jimmunol.1600584
Nandi, B., & Behar, S. M. (2011). Regulation of neutrophils by interferon-γ limits lung inflammation during tuberculosis infection. *The Journal of Experimental Medicine*, *208*(11), 2251–2262. https://doi.org/10.1084/jem.20110919
Nathan, C., & Shiloh, M. U. (2000). Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(16), 8841–8848. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10922044
Nayak, S., & Acharjya, B. (2012). Mantoux test and its interpretation. *Indian Dermatology Online Journal*, *3*(1), 2–6. https://doi.org/10.4103/2229-5178.93479
Nguyen, D. T., Teeter, L. D., Graves, J., & Graviss, E. A. (2018). Characteristics Associated with Negative Interferon-γ Release Assay Results in Culture-Confirmed Tuberculosis Patients, Texas, USA, 2013–2015. *Emerging Infectious Diseases*, *24*(3), 534–540. https://doi.org/10.3201/eid2403.171633
Nieuwenhuizen, N. E., & Kaufmann, S. H. E. (2018). Next-Generation Vaccines Based on Bacille Calmette–Guérin. *Frontiers in Immunology*, *9*, 121. https://doi.org/10.3389/fimmu.2018.00121
North, R. J., & Jung, Y.-J. (2004). I mmunity to T uberculosis. *Annual Review of Immunology*, *22*(1), 599–623. https://doi.org/10.1146/annurev.immunol.22.012703.104635
O'Garra, A., Redford, P. S., McNab, F. W., Bloom, C. I., Wilkinson, R. J., & Berry, M. P. R. (2013). The Immune Response in Tuberculosis. *Annual Review of Immunology*, *31*(1), 475–527. https://doi.org/10.1146/annurev-immunol-032712-095939
Ordway, D., Henao-Tamayo, M., Harton, M., Palanisamy, G., Troudt, J., Shanley, C., … Orme, I. M. (2007). The hypervirulent Mycobacterium tuberculosis strain HN878 induces a potent TH1 response followed by rapid down-regulation. *Journal of Immunology (Baltimore, Md. : 1950)*, *179*(1), 522–531. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17579073
Oshima, K., Haeger, S. M., Hippensteel, J. A., Herson, P. S., & Schmidt, E. P. (2018). More than a biomarker: the systemic consequences of heparan sulfate fragments released during endothelial surface layer degradation (2017 Grover Conference Series). *Pulmonary Circulation*, *8*(1), 2045893217745786. https://doi.org/10.1177/2045893217745786
Ottenhoff, T. H. M., Dass, R. H., Yang, N., Zhang, M. M., Wong, H. E. E., Sahiratmadja, E., … Hibberd, M. L. (2012). Genome-Wide Expression Profiling Identifies Type 1 Interferon Response Pathways in Active Tuberculosis. *PLoS ONE*, *7*(9), e45839. https://doi.org/10.1371/journal.pone.0045839
Palmero, D., Eiguchi, K., Rendo, P., Castro Zorrilla, L., Abbate, E., & González Montaner, L. J. (1999). Phase II trial of recombinant interferon-α2b in patients with advanced intractable multidrug-resistant pulmonary tuberculosis: long-term follow-up. *Int J Tuberc Lung Dis*, *3*(3), 214–218.
Pan, H., Yan, B.-S., Rojas, M., Shebzukhov, Y. V., Zhou, H., Kobzik, L., … Kramnik, I. (2005). Ipr1 gene mediates innate immunity to tuberculosis. *Nature*, *434*(7034), 767–772. https://doi.org/10.1038/nature03419
Paone, J. F., Waalkes, T. P., Baker, R. R., & Shaper, J. H. (1980). Serum UDP-galactosyl transferase as a potential biomarker for breast carcinoma. *Journal of Surgical Oncology*, *15*(1), 59–66. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6775160

Pascual, V., Chaussabel, D., & Banchereau, J. (2010). A genomic approach to human autoimmune diseases. *Annual Review of Immunology*, *28*, 535–571. https://doi.org/10.1146/annurev-immunol-030409-101221

Pearl, J. E., Saunders, B., Ehlers, S., Orme, I. M., & Cooper, A. M. (2001). Inflammation and Lymphocyte Activation during Mycobacterial Infection in the Interferon-γ-Deficient Mouse. *Cellular Immunology*, *211*(1), 43–50. https://doi.org/10.1006/cimm.2001.1819

Perkins, M. D. (2009). New diagnostics for tuberculosis. In H. S. Schaaf, A. I. Zumla, J. M. Grange, M. C. Raviglione, W. W. Yew, J. R. Starke, … P. R. Donald (Eds.), *Tuberculosis. A Comprehensive Clinical Reference* (pp. 227–236). Edinburgh: W.B. Saunders. https://doi.org/https://doi.org/10.1016/B978-1-4160-3988-4.00023-8

Pestka, S., Krause, C. D., & Walter, M. R. (2004). Interferons, interferon-like cytokines, and their receptors. *Immunological Reviews*, *202*(1), 8–32. https://doi.org/10.1111/j.0105-2896.2004.00204.x

Phuah, J. Y., Mattila, J. T., Lin, P. L., & Flynn, J. L. (2012). Activated B Cells in the Granulomas of Nonhuman Primates Infected with Mycobacterium tuberculosis. *The American Journal of Pathology*, *181*(2), 508–514. https://doi.org/10.1016/j.ajpath.2012.05.009

Platanias, L. C. (2005). Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nature Reviews Immunology*, *5*(5), 375–386. https://doi.org/10.1038/nri1604

Priya, G. B., Nagaleekar, V. K., Milton, A. A. P., Saminathan, M., Kumar, A., Sahoo, A. R., … Gandham, R. K. (2017). Genome wide host gene expression analysis in mice experimentally infected with Pasteurella multocida. *PLOS ONE*, *12*(7), e0179420. https://doi.org/10.1371/journal.pone.0179420

R Core Team, R. (2018). R: A Language and Environment for Statistical Computing. (R. D. C. Team, Ed.), *R Foundation for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://doi.org/10.1007/978-3-540-74686-7

Reiley, W. W., Calayag, M. D., Wittmer, S. T., Huntington, J. L., Pearl, J. E., Fountain, J. J., … Woodland, D. L. (2008). ESAT-6-specific CD4 T cell responses to aerosol Mycobacterium tuberculosis infection are initiated in the mediastinal lymph nodes. *Proceedings of the National Academy of Sciences*, *105*(31), 10961–10966. https://doi.org/10.1073/pnas.0801496105

Riley, R. L., & O'Grady, F. (1961). *Airborne infection: transmission and control.* Macmillan. Retrieved from https://books.google.de/books/about/Airborne_infection_transmission_and_cont.html?id=qztrAAAAMAAJ&redir_esc=y

Risso, A. (2000). Leukocyte antimicrobial peptides: multifunctional effector molecules of innate immunity. *Journal of Leukocyte Biology*, *68*(6), 785–792. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11129645

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Rusinova, I., Forster, S., Yu, S., Kannan, A., Masse, M., Cumming, H., … Hertzog, P. J. (2012). INTERFEROME v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Research*, *41*(D1), D1040–D1046. https://doi.org/10.1093/nar/gks1215

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Saito, M., Iwawaki, T., Taya, C., Yonekawa, H., Noda, M., Inui, Y., … Kohno, K. (2001). Diphtheria toxin receptor–mediated conditional and targeted cell ablation in transgenic mice. *Nature Biotechnology*, *19*(8), 746–750. https://doi.org/10.1038/90795

Sambarey, A., Devaprasad, A., Baloni, P., Mishra, M., Mohan, A., Tyagi, P., … Chandra, N. (2017). Meta-analysis of host response networks identifies a common core in tuberculosis. *Npj Systems Biology and Applications*, *3*(4). https://doi.org/10.1038/s41540-017-0005-4

Sánchez, F., Radaeva, T. V, Nikonenko, B. V, Persson, A.-S., Sengul, S., Schalling, M., … Lavebratt,

C. (2003). Multigenic control of disease severity after virulent Mycobacterium tuberculosis infection in mice. *Infection and Immunity*, *71*(1), 126–131. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12496157

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, *26*(5), 1651–1686. https://doi.org/10.1214/aos/1024691352

Schlesinger, L. S. (1996). Entry of Mycobacterium tuberculosis into mononuclear phagocytes. *Current Topics in Microbiology and Immunology*, *215*, 71–96. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8791710

Schouten, M., Daan de Boer, J., Kager, L. M., Roelofs, J. J. T. H., Meijers, J. C. M., Esmon, C. T., … van der Poll, T. (2014). The endothelial protein C receptor impairs the antibacterial response in murine pneumococcal pneumonia and sepsis. *Thrombosis and Haemostasis*, *111*(05), 970–980. https://doi.org/10.1160/TH13-10-0859

Schulze, A., & Downward, J. (2001). Navigating gene expression using microarrays — a technology review. *Nature Cell Biology*, *3*(8), E190–E195. https://doi.org/10.1038/35087138

Seok, J., Warren, H. S., Cuenca, A. G., Mindrinos, M. N., Baker, H. V, Xu, W., … Tompkins, R. G. (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(9), 3507–3512. https://doi.org/10.1073/pnas.1222878110

Sharma, S. K., Mohan, A., & Sharma, A. (2016). Miliary tuberculosis: A new look at an old foe. *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, *3*, 13–27. https://doi.org/10.1016/J.JCTUBE.2016.03.003

Shay, T., Jojic, V., Zuk, O., Rothamel, K., Puyraimond-Zemmour, D., Feng, T., … Regev, A. (2013). Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(8), 2946–2951. https://doi.org/10.1073/pnas.1222738110

Shlens, J. (2014). A Tutorial on Principal Component Analysis. Retrieved from http://arxiv.org/abs/1404.1100

Simon, R. M., Korn, E., McShane, L., Wright, G., & Zhao, Y. (2003). DNA Microarray Technology. In *Design and analysis of DNA microarray investigations* (p. 199). Bethesda: Springer.

Skeen, M. J., & Ziegler, H. K. (2018). IL-1 and IL-12. bacteria via macrophage-derived cytokines production of IFN-gamma is mediated by Activation of gamma delta T cells for. Retrieved from http://www.jimmunol.org/content/154/11/5832

Srivastava, S., & Ernst, J. D. (2013). Cutting Edge: Direct Recognition of Infected Cells by CD4 T Cells Is Required for Control of Intracellular Mycobacterium tuberculosis In Vivo. *The Journal of Immunology*, *191*(3), 1016–1020. https://doi.org/10.4049/jimmunol.1301236

Stanton, L. W. (2001). Methods to profile gene expression. *Trends in Cardiovascular Medicine*, *11*(2), 49–54. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11530292

Starke, J. R. (1996). Tuberculosis Skin Testing: New Schools of Thought. *Pediatrics*, *98*(1). Retrieved from http://pediatrics.aappublications.org/content/98/1/123

Strutt, T. M., McKinstry, K. K., Dibble, J. P., Winchell, C., Kuang, Y., Curtis, J. D., … Swain, S. L. (2010). Memory CD4+ T cells induce innate responses independently of pathogen. *Nature Medicine*, *16*(5), 558–564. https://doi.org/10.1038/nm.2142

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Suliman, S., Thompson, E. G., Sutherland, J., Weiner, J., Ota, M. O. C., Shankar, S., … Geiter, L. (2018). Four-Gene Pan-African Blood Signature Predicts Progression to Tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, *197*(9), 1198–1208. https://doi.org/10.1164/rccm.201711-2340OC

Sunderkotter, C., Nikolic, T., Dillon, M. J., van Rooijen, N., Stehling, M., Drevets, D. A., & Leenen, P. J. M. (2004). Subpopulations of Mouse Blood Monocytes Differ in Maturation Stage and Inflammatory Response. *The Journal of Immunology*, *172*(7), 4410–4417.

https://doi.org/10.4049/jimmunol.172.7.4410
Sutherland, A., Thomas, M., Brandon, R. A., Brandon, R. B., Lipman, J., Tang, B., … Venter, D. (2011). Development and validation of a novel molecular biomarker diagnostic test for the early detection of sepsis. *Critical Care*, *15*(3), R149. https://doi.org/10.1186/cc10274
Sweeney, T. E., Braviak, L., Tato, C. M., & Khatri, P. (2016). Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. *The Lancet Respiratory Medicine*, *4*(3), 213–224. https://doi.org/10.1016/S2213-2600(16)00048-5
Takao, K., & Miyakawa, T. (2014). Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences*, 1–6. https://doi.org/10.1073/pnas.1401965111
Tang, B. M. P., McLean, A. S., Dawes, I. W., Huang, S. J., & Lin, R. C. Y. (2009). Gene-expression profiling of peripheral blood mononuclear cells in sepsis. *Critical Care Medicine*, *37*(3), 882–888. https://doi.org/10.1097/CCM.0b013e31819b52fd
The Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, *45*(D1), D331–D338. https://doi.org/10.1093/nar/gkw1108
Thuong, N. T. T., Dunstan, S. J., Chau, T. T. H., Thorsson, V., Simmons, C. P., Quyen, N. T. H., … Hawn, T. R. (2008). Identification of Tuberculosis Susceptibility Genes with Human Macrophage Gene Expression Profiles. *PLoS Pathogens*, *4*(12), e1000229. https://doi.org/10.1371/journal.ppat.1000229
Tientcheu, L. D., Maertzdorf, J., Weiner, J., Adetifa, I. M., Mollenkopf, H.-J., Sutherland, J. S., … Ota, M. O. (2015). Differential transcriptomic and metabolic profiles of M. africanum- and M. tuberculosis-infected patients after, but not before, drug treatment. *Genes & Immunity*, *16*(5), 347–355. https://doi.org/10.1038/gene.2015.21
Trinath, J., Maddur, M. S., Kaveri, S. V., Balaji, K. N., & Bayry, J. (2012). Mycobacterium tuberculosis Promotes Regulatory T-Cell Expansion via Induction of Programmed Death-1 Ligand 1 (PD-L1, CD274) on Dendritic Cells. *The Journal of Infectious Diseases*, *205*(4), 694–696. https://doi.org/10.1093/infdis/jir820
Turner, J., Gonzalez-Juarrero, M., Saunders, B. M., Brooks, J. V., Marietta, P., Ellis, D. L., … Orme, I. M. (2001). Immunological Basis for Reactivation of Tuberculosis in Mice. *Infection and Immunity*, *69*(5), 3264–3270. https://doi.org/10.1128/IAI.69.5.3264-3270.2001
Ulrichs, T., Kosmiadi, G. A., Trusov, V., Jörg, S., Pradl, L., Titukhina, M., … Kaufmann, S. H. (2004). Human tuberculous granulomas induce peripheral lymphoid follicle-like structures to orchestrate local host defence in the lung. *The Journal of Pathology*, *204*(2), 217–228. https://doi.org/10.1002/path.1628
Verhagen, L. M., Zomer, A., Maes, M., Villalba, J. A., del Nogal, B., Eleveld, M., … Hermans, P. W. (2013). A predictive signature gene set for discriminating active from latent tuberculosis in Warao Amerindian children. *BMC Genomics*, *14*(1), 74. https://doi.org/10.1186/1471-2164-14-74
Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., & Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, *19*(2), 327–335. https://doi.org/10.1101/gr.073585.107
Walter, N. D., Miller, M. A., Vasquez, J., Weiner, M., Chapman, A., Engle, M., … Geraci, M. W. (2016). Blood Transcriptional Biomarkers for Active Tuberculosis among Patients in the United States: a Case-Control Study with Systematic Cross-Classifier Evaluation. *Journal of Clinical Microbiology*, *54*(2), 274–282. https://doi.org/10.1128/JCM.01990-15
Wan-Chung Hu, B. (2013). Sepsis is a syndrome with hyperactivity of TH17-like innate immunity and hypoactivity of adaptive immunity. *ArXiv.Org*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1311/1311.4747.pdf
Ward, C. M., Jyonouchi, H., Kotenko, S. V., Smirnov, S. V., Patel, R., Aguila, H., … Holland, S. M. (2007). Adjunctive treatment of disseminated Mycobacterium avium complex infection with interferon alpha-2b in a patient with complete interferon-gamma receptor R1 deficiency. *European Journal of Pediatrics*, *166*(9), 981–985. https://doi.org/10.1007/s00431-006-0339-1
Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845
Weiner, J. (2017). tmod: Module enrichment tool. Retrieved from http://bioinfo.mpiib-

berlin.mpg.de/tmod/
Weiner, J. 3rd, & Domaszewska, T. (2016). tmod: an R package for general and multivariate enrichment analysis. *PeerJ Preprints*, *No. e2420v*. https://doi.org/10.7287/PEERJ.PREPRINTS.2420V1
Wells, W. F. (1934). On air-borne infection. Study II. Droplets and droplet nuclei. *American Journal of Hygiene*, *20*, 611–618. Retrieved from https://academic.oup.com/aje/article-abstract/20/3/611/280025
WHO. (2008). *WHO policy statement: molecular line probe assays for rapid screening of patients at risk of multidrug-resistant tuberculosis*. *WHO*. World Health Organization. Retrieved from http://www.who.int/tb/laboratory/line_probe_assays/en/
WHO. (2011). *Commercial serodiagnostic tests for diagnosis of tuberculosis : policy statement.*
WHO. (2013a). *Guideline : Nutritional care and support for patients with tuberculosis*. *World Health Organization*. Retrieved from http://apps.who.int/iris/bitstream/10665/94836/1/9789241506410_eng.pdf
WHO. (2013b). *Xpert MTB/RIF assay for the diagnosis of pulmonary and extrapulmonary TB in adults and children*. *WHO*. World Health Organization. Retrieved from http://www.who.int/tb/publications/xpert-mtb-rif-assay-diagnosis-policy-update/en/
WHO. (2014). *Companion handbook to the WHO guidelines for the programmatic management of drug-resistant tuberculosis.*
WHO. (2015a). *Factsheet: Post-2015 Global TB Strategy and targets*. *WHO*. World Health Organization. Retrieved from http://www.who.int/tb/post2015_strategy/en/
WHO. (2015b). *Systematic screening for active tuberculosis: principles and recommendations*. *WHO*. World Health Organization. Retrieved from http://www.who.int/tb/tbscreening/en/
WHO. (2017). *Global tuberculosis report 2017*. *WHO*. World Health Organization. Retrieved from http://www.who.int/tb/publications/global_report/en/
Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.
Winslow, G. M., Roberts, A. D., Blackman, M. A., & Woodland, D. L. (2003). Persistence and turnover of antigen-specific CD4 T cells during chronic tuberculosis infection in the mouse. *Journal of Immunology (Baltimore, Md. : 1950)*, *170*(4), 2046–2052. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12574375
Wong, H. R., Cvijanovich, N., Allen, G. L., Lin, R., Anas, N., Meyer, K., … Genomics of Pediatric SIRS/Septic Shock Investigators. (2009). Genomic expression profiling across the pediatric systemic inflammatory response syndrome, sepsis, and septic shock spectrum. *Critical Care Medicine*, *37*(5), 1558–1566. https://doi.org/10.1097/CCM.0b013e31819fcc08
Wu, L. S.-H., Lee, S.-W., Huang, K.-Y., Lee, T.-Y., Hsu, P. W.-C., & Weng, J. T.-Y. (2014). Systematic expression profiling analysis identifies specific microRNA-gene interactions that may differentiate between active and latent tuberculosis infection. *BioMed Research International*, *2014*, 895179. https://doi.org/10.1155/2014/895179
Yamaguchi, K. D., Ruderman, D. L., Croze, E., Wagner, T. C., Velichko, S., Reder, A. T., & Salamon, H. (2008). IFN-beta-regulated genes show abnormal expression in therapy-naïve relapsing-remitting MS mononuclear cells: gene expression analysis employing all reported protein-protein interactions. *Journal of Neuroimmunology*, *195*(1–2), 116–120. https://doi.org/10.1016/j.jneuroim.2007.12.007
Zak, D. E., Penn-Nicholson, A., Scriba, T. J., Thompson, E., Suliman, S., Amon, L. M., … Hanekom, W. A. (2016). A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *The Lancet*. https://doi.org/10.1016/S0140-6736(15)01316-1
Zarogoulidis, P., Kioumis, I., Papanas, N., Manika, K., Kontakiotis, T., Papagianis, A., & Zarogoulidis, K. (2012). The effect of combination IFN-alpha-2a with usual antituberculosis chemotherapy in non-responding tuberculosis and diabetes mellitus: a case report and review of the literature. *Journal of Chemotherapy*, *24*(3), 173–177. https://doi.org/10.1179/1973947812Y.0000000005
Zhang, P., Li, Y., Zhang, L.-D., Wang, L.-H., Wang, X., He, C., & Lin, Z.-F. (2014). Proteome changes in mesenteric lymph induced by sepsis. *Molecular Medicine Reports*, *10*(6), 2793–2804. https://doi.org/10.3892/mmr.2014.2580
Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *3*(1), 1–130.

https://doi.org/10.2200/S00196ED1V01Y200906AIM006
Zuñiga, J., Torres-García, D., Santos-Mendoza, T., Rodriguez-Reyna, T. S., Granados, J., & Yunis, E. J. (2012). Cellular and Humoral Mechanisms Involved in the Control of Tuberculosis. *Clinical and Developmental Immunology*, *2012*, 1–18. https://doi.org/10.1155/2012/193923

8. SUPPLEMENTARY MATERIAL

Supplementary Table 1. Assignment of the genes from the original Li et al. and Chaussabel et al. modules to the new IFN type I, IFN type II and IFN type I and II modules

Original ID and name are derived from (Chaussabel et al., 2008; Li et al., 2014). IFN type is based on the assignment by Interferome database (Rusinova et al., 2012).

Original ID	Original name	genes	IFN type	New module	New ID
LI.M127	type I IFN response	TAP1	I and II	Type I and II IFN module 1	LI.M127_IandII
		IFIH1			
		IRF7	I and II	Type I and II IFN module 1	LI.M127_IandII
		PARP9	I and II	Type I and II IFN module 1	LI.M127_IandII
		STAT1	I and II	Type I and II IFN module 1	LI.M127_IandII
		PLSCR1	I and II	Type I and II IFN module 1	LI.M127_IandII
		IFITM1	I and II	Type I and II IFN module 1	LI.M127_IandII
		HERC5	I		
		DDX60	I and II	Type I and II IFN module 1	LI.M127_IandII
		USP18	I		
		RSAD2	I and II	Type I and II IFN module 1	LI.M127_IandII
		IFIT1	I and II	Type I and II IFN module 1	LI.M127_IandII
LI.M75	Antiviral IFN signature	IFIH1			
		ELANE			
		SERPING1	I and II	Type I and II IFN module 2	LI.M75_IandII
		IL1B	I	Type I IFN module 1	LI.M75_I
		RSAD2	I and II	Type I and II IFN module 2	LI.M75_IandII
		IFIT1	I and II	Type I and II IFN module 2	LI.M75_IandII
		RARA			
		DDX58	I	Type I IFN module 1	LI.M75_I
		FCER1A	I	Type I IFN module 1	LI.M75_I
		DHX58	I and II	Type I and II IFN module 2	LI.M75_IandII
		PTX3			
		CARD9			
		OAS1	I and II	Type I and II IFN module 2	LI.M75_IandII
		OAS3	I	Type I IFN module 1	LI.M75_I
		PML	I and II	Type I and II IFN module 2	LI.M75_IandII
		ANXA3			
		HERC5	I	Type I IFN module 1	LI.M75_I
		DDX60	I and II	Type I and II IFN module 2	LI.M75_IandII
		CXCL10	I	Type I IFN module 1	LI.M75_I
		IRF7	I and II	Type I and II IFN module 2	LI.M75_IandII
		C1QB			
		BCL3			
LI.M158.0	IFN alpha response (I)	LHCGR			
		COL8A1			
		IMPG2			
		ITGB4			
		MMP12	I		
		TNR			
		IFNA7			
		IFNA4			
		SFN			
		LAMC2			
		ST14			
		ADAMTS20			
		FGF5			
		IFNA10	I		

		IFNA16			
		IFNA14			
LI.M158.1	IFN alpha response (II)	LHCGR			
		AMER2			
		IFNA8			
		IFNA2			
		IFNA7			
		IFNA5			
		IFNA4			
		IFNA21			
		PRL			
		ADAMTS20			
		IFNA10	I		
		IFNA16			
		IFNA14			
DC.M5.12		RBCK1	I	Type I IFN module 2	DC.M5.12_I
		TRAFD1	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		TRIM21	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		LOC401433			
		RFWD2			
		CHM5			
		TAP2	I	Type I IFN module 2	DC.M5.12_I
		SP110	I	Type I IFN module 2	DC.M5.12_I
		GADD45B			
		IFI16	I	Type I IFN module 2	DC.M5.12_I
		TAP1	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		ZNFX1	I	Type I IFN module 2	DC.M5.12_I
		PHF11	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		ACTA2	I	Type I IFN module 2	DC.M5.12_I
		C1QA	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		SP140	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		ABCA1			
		TCN2	I nad II	Type I and II IFN module 3	DC.M5.12_IandII
		ZC3HAV1	I	Type I IFN module 2	DC.M5.12_I
		HSH2D	I	Type I IFN module 2	DC.M5.12_I
		LOC554203			
		GBP2	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		TRIM5	I	Type I IFN module 2	DC.M5.12_I
		RHBDF2			
		TMEM140	I	Type I IFN module 2	DC.M5.12_I
		ADAR	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		BTN3A1	II		
		PARP10	I nad II	Type I and II IFN module 3	DC.M5.12_IandII
		LGALS9	I	Type I IFN module 2	DC.M5.12_I
		NBN	I	Type I IFN module 2	DC.M5.12_I
		TYMP	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		SAMD9	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		SRBD1			
		NCOA7	I	Type I IFN module 2	DC.M5.12_I
		DRAP1	I	Type I IFN module 2	DC.M5.12_I
		UNC93B1	I	Type I IFN module 2	DC.M5.12_I
		SP100	I	Type I IFN module 2	DC.M5.12_I
		NTNG2	I	Type I IFN module 2	DC.M5.12_I
		DHRS9	II		
		TDRD7	I	Type I IFN module 2	DC.M5.12_I
		TRANK1	I	Type I IFN module 2	DC.M5.12_I
		MDK	I	Type I IFN module 2	DC.M5.12_I
		NT5C3A	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		ASPRV1			

		IRF9	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		REC8			
		RNF213	I	Type I IFN module 2	DC.M5.12_I
		ISG20	I nad II	Type I and II IFN module 3	DC.M5.12_IandII
		DYNLT1	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		LHFPL2			
		TRIM56	I	Type I IFN module 2	DC.M5.12_I
		TRIM25	I	Type I IFN module 2	DC.M5.12_I
		TRIM38	I	Type I IFN module 2	DC.M5.12_I
		ETV7	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		PSMB9	I and II	Type I and II IFN module 3	DC.M5.12_IandII
		CPT1B			
		BST2	I	Type I IFN module 2	DC.M5.12_I
		CASP1	I	Type I IFN module 2	DC.M5.12_I
		NMI	I and II	Type I and II IFN module 3	DC.M5.12_IandII
DC.M1.2	IFN	LY6E	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		IFIT1	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		OAS1	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		IFIT3	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		OAS3	I	Type I IFN module 3	DC.M1.2_I
		OASL	I	Type I IFN module 3	DC.M1.2_I
		LOC129607			
		ISG15	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		HERC5	I	Type I IFN module 3	DC.M1.2_I
		MX1	I	Type I IFN module 3	DC.M1.2_I
		BATF2	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		LAMP3	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		IFI44L	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		XAF1	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		IFI44	I	Type I IFN module 3	DC.M1.2_I
		OAS2	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		TRIM6			
		HES4	I	Type I IFN module 3	DC.M1.2_I
		OTOF	I	Type I IFN module 3	DC.M1.2_I
		FLJ20035			
		IFITM3	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		CXCL10	I	Type I IFN module 3	DC.M1.2_I
		EPSTI1	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		SERPING1	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		LOC26010			
		RSAD2	I and II	Type I and II IFN module 4	DC.M1.2_IandII
		RTP4	I and II	Type I and II IFN module 4	DC.M1.2_IandII
DC.M3.4	IFN	IFIH1			
		IRF7	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		PARP14	I	Type I IFN module 4	DC.M3.4_I
		IFIT2	I	Type I IFN module 4	DC.M3.4_I
		IFI35	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		SAMD9L	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		STAT1	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		OAS2	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		IFIT5	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		ATF3	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		geneSEPT4	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		HERC6	I	Type I IFN module 4	DC.M3.4_I
		IFITM1	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		TRIM78P		Type I IFN module 4	
		EIF2AK2	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		AIM2	I and II	Type I and II IFN module 5	DC.M3.4_IandII
		MT1A			
		MOV10	I and II	Type I and II IFN module 5	DC.M3.4_IandII

CCL8	I and II	Type I and II IFN module 5	DC.M3.4_IandII
HELZ2	I	Type I IFN module 4	DC.M3.4_I
ZBP1	I and II	Type I and II IFN module 5	DC.M3.4_IandII
WARS	I and II	Type I and II IFN module 5	DC.M3.4_IandII
LAP3	I and II	Type I and II IFN module 5	DC.M3.4_IandII
GBP5	I and II	Type I and II IFN module 5	DC.M3.4_IandII
TNFSF10	I and II	Type I and II IFN module 5	DC.M3.4_IandII
GBP1			
FBXO6	I and II	Type I and II IFN module 5	DC.M3.4_IandII
PARP10	I and II	Type I and II IFN module 5	DC.M3.4_IandII
TRIM22	I and II	Type I and II IFN module 5	DC.M3.4_IandII
GBP3	I	Type I IFN module 4	DC.M3.4_I
ZNF684			
CARD17			
GALM	I	Type I IFN module 4	DC.M3.4_I
DHX58	I and II	Type I and II IFN module 5	DC.M3.4_IandII
CEACAM 1			
UBE2L6	I and II	Type I and II IFN module 5	DC.M3.4_IandII
PML	I and II	Type I and II IFN module 5	DC.M3.4_IandII
APOL6	I and II	Type I and II IFN module 5	DC.M3.4_IandII
SOCS1	I and II	Type I and II IFN module 5	DC.M3.4_IandII
LGALS3BP	I and II	Type I and II IFN module 5	DC.M3.4_IandII
SCO2			
DDX58	I	Type I IFN module 4	DC.M3.4_I
TNFAIP6	I and II	Type I and II IFN module 5	DC.M3.4_IandII
IDO1	I and II	Type I and II IFN module 5	DC.M3.4_IandII
MT2A			
GBP6			
STAT2	I and II	Type I and II IFN module 5	DC.M3.4_IandII
TIMM10			
PARP12	I and II	Type I and II IFN module 5	DC.M3.4_IandII
PLSCR1	I and II	Type I and II IFN module 5	DC.M3.4_IandII
PARP9	I and II	Type I and II IFN module 5	DC.M3.4_IandII
LOC400759			
GBP4	I and II	Type I and II IFN module 5	DC.M3.4_IandII

Supplementary Table 2 IFN type I signaling genes present in MDS according to Interferome database classification

1	A2M	AAK1	ABCA6	ABCC3	ABCD1	ABCD2	ABCD3
8	ABCF1	ABCG2	ABHD15	ABHD2	ABHD3	ABL2	ABTB2
15	ACACA	ACAD9	ACO1	ACOT9	ACOX1	ACTA2	ACVR1
22	ACYP1	ADA	ADAM12	ADAMDEC1	ADAP2	ADD3	ADPRH
29	ADPRHL2	ADRB2	ADSL	AFF1	AGO1	AGRN	AHR
36	AKAP10	AKR1C3	ALAD	ALDH1A2	ALDH3A2	ALDH5A1	ALKBH1
43	ALOX5AP	AMOTL2	AMPD2	AMT	AMY1C	ANAPC5	ANGEL2
50	ANGPT2	ANGPTL4	ANK2	ANK3	ANKFY1	ANKRD12	ANKRD55
57	ANKS1A	ANXA2R	AOC1	AP1S2	AP3S2	APEX1	APLNR
64	APLP2	APOBEC3A	APOBEC3B	APOBEC3F	APOC1	APOC2	APOC4-APOC2
71	APOE	APOM	ARAP1	ARAP2	ARF3	ARHGAP19	ARHGAP23
78	ARHGAP25	ARHGAP27	ARHGAP29	ARHGAP33	ARHGAP35	ARHGEF15	ARHGEF2
85	ARHGEF3	ARHGEF9	ARID3A	ARID5A	ARL15	ARL2	ARL2BP
92	ARL4C	ARMCX1	ARRB1	ARRDC2	ARSD	ARSE	ASAH2
99	ASAP2	ASB1	ASB14	ASGR1	ASS1	ATF1	ATG10
106	ATG14	ATG2A	ATL3	ATM	ATP1B2	ATP2B1	ATP2B4
113	ATP5B	ATP6AP2	ATP6V0A2	ATP6V0B	ATP6V1G2-DDX39B	ATPAF2	ATRIP
120	ATRN	ATRX	ATXN7	AUTS2	AXIN2	AZI2	B3GNT2
127	BACH1	BAG1	BARD1	BATF3	BAX	BBS4	BCKDHA
134	BCL6	BCL7A	BCS1L	BICD2	BLNK	BLVRA	BNIP1
141	BPGM	BRD3	BRD8	BRI3BP	BRIP1	BST2	BTG1
148	BTG2	BUB1	C11orf58	C12orf57	C12orf76	C14orf28	C17orf75
155	C19orf66	C1GALT1	C1orf27	C21orf91	C2CD2	C2orf27A	C4BPA
162	C4orf33	C4orf46	C6orf48	C9orf66	CACNA1A	CACNG1	CACTIN
169	CALCRL	CALML3	CAMK1	CAMK2G	CAMSAP2	CAMTA1	CAMTA2
176	CANX	CARD16	CARF	CASP1	CASP10	CASP2	CASP4
183	CASP7	CBFA2T3	CBR1	CBWD1	CBWD2	CBWD3	CBX1
190	CBX3	CBX5	CCDC62	CCL11	CCL2	CCL20	CCL22
197	CCL24	CCNA1	CCNE1	CCNG1	CCNG2	CCR1	CCRL2
204	CD101	CD14	CD163	CD164	CD207	CD2AP	CD2BP2
211	CD68	CD69	CD79B	CD80	CD86	CD99	CDC42EP3
218	CDH18	CDK17	CDK5	CDK5R1	CDK6	CDKL2	CDKN2C
225	CDKN2D	CEBPA	CENPF	CERK	CERS5	CETN3	CFD
232	CH25H	CHAF1A	CHN2	CHRNB1	CHRNE	CHST15	CHSY1
239	CITED2	CLDN1	CLDN23	CLEC10A	CLEC3B	CLEC4C	CLECL1
246	CLMN	CLOCK	CLUH	CMKLR1	CMTR1	CNDP2	CNN2
253	CNPPD1	COL7A1	COMMD1	COQ10A	CORO1B	COX7A1	CPD
260	CPEB4	CPED1	CPM	CPNE3	CREB3L2	CREBL2	CROT
267	CRYAB	CSDE1	CSNK1G1	CSRNP1	CSRP2	CST6	CST7
274	CSTA	CSTF1	CSTF3	CTAGE5	CTDNEP1	CTNNAL1	CX3CR1
281	CXCL1	CXCL10	CXCL12	CXCL16	CXCL2	CXCL3	CXCL5
288	CXCR2	CXCR4	CXorf21	CYB5R1	CYBRD1	CYP11B1	CYP11B2

295	CYP27A1	CYP2J2	CYP3A7	CYTH2	DAPK1	DBF4	DBF4B
302	DCP2	DDB2	DDIT3	DDO	DDX28	DDX39B	DDX58
309	DEDD2	DESI1	DFNA5	DGKG	DGKZ	DHCR24	DHCR7
316	DHFR	DHRS3	DHX57	DIAPH2	DLG4	DLGAP2	DMTF1
323	DNAAF1	DNAJB2	DNAJC25	DNAJC25-GNG10	DOK1	DOK2	DPH2
330	DRAP1	DSP	DSTYK	DTL	DTX3L	DUSP22	DUSP4
337	DUSP5	DUT	DYNC2LI1	DYSF	E2F5	EBNA1BP2	EBP
344	ECE1	EEF1B2	EEF1E1	EGR1	EHD4	EIF2B2	EIF3C
351	EIF3D	EIF3L	EIF4G3	ELF4	ELK4	ELL2	ELMSAN1
358	ELOVL5	EML3	ENDOD1	ENG	ENOSF1	ENTPD1	EPDR1
365	EPHB2	ERC2	ERCC2	ERF	ETS2	ETV6	EVI2A
372	EVI2B	EXO5	EXOC3L1	EXOSC2	EXOSC7	EXOSC9	EXT1
379	EXT2	EZH1	F3	FADS1	FADS2	FAM102A	FAM110A
386	FAM122C	FAM126A	FAM13A	FAM167A	FAM168B	FAM177A1	FAM20B
393	FAM216A	FAM46A	FAM47E-STBD1	FAM49A	FAM50B	FAM58A	FAM72A
400	FAM72C	FAM72D	FAM76A	FAM84B	FARP2	FASN	FBRSL1
407	FBXO46	FBXW2	FCAR	FCER1A	FCGR2B	FCGRT	FCHO1
414	FCRL3	FDFT1	FDPS	FDXR	FFAR2	FGD2	FGD6
421	FGF2	FGFBP2	FIG4	FN1	FOLH1	FOS	FOXC1
428	FOXJ1	FOXJ2	FOXO1	FOXO3	FPR3	FRAT2	FRMD3
435	FRMD4B	FUT4	FUT8	FXYD6	FYN	FZD1	FZD3
442	GABARAPL1	GADD45A	GAK	GAL3ST4	GALM	GALNT1	GART
449	GBE1	GBP3	GCA	GCLM	GGA2	GGT5	GINS2
456	GK	GK2	GLCE	GLRX	GLT8D1	GLTP	GMDS
463	GMPR	GNA11	GNA12	GNA15	GNAI3	GNAT1	GNB4
470	GNG10	GNPDA1	GNS	GOLM1	GPBAR1	GPD2	GPI
477	GPN1	GPNMB	GPR141	GPR15	GPR161	GPR18	GPR180
484	GPR183	GPR3	GPR65	GPR68	GPX3	GRAMD4	GRIK2
491	GRM1	GTF2E2	GTF2H3	GTPBP8	GUCD1	GUCY1A2	GYPA
498	H1F0	H1FX	H2AFV	H2BFS	HAGH	HAVCR2	HCAR2
505	HCAR3	HDAC5	HDHD2	HDX	HEG1	HELB	HELZ2
512	HERC5	HERC6	HES4	HESX1	HGF	HINT1	HINT2
519	HINT3	HIP1	HIPK2	HIRA	HIRIP3	HIST1H1E	HIST1H2AC
526	HIST1H2BD	HIST1H3I	HIST1H4C	HIST2H2AA3	HIST2H2AA4	HIST2H2AC	HIST2H2BE
533	HK1	HLA-J	HLF	HMBS	HMGB2	HMGN2P46	HMGN3
540	HMOX1	HNMT	HNRNPA1	HNRNPA1P10	HNRNPM	HOMER1	HOXB2
547	HOXB6	HPSE	HRK	HSF2	HSH2D	HSPA4	ICA1
554	ICOSLG	ID2	IDS	IDUA	IER3	IER5	IFI16
561	IFI27	IFI44	IFIT2	IFNA10	IFNA17	IFNGR2	IFRD1
568	IGF1R	IGF2BP3	IGFBP4	IL18	IL1A	IL1B	IL23A
575	IL27	IL27RA	IL4I1	IL7	IMPA2	IMPDH2	INE1
582	INHBA	INPP5J	INSIG1	IPO11	IPO5	IPO9	IRAK2
589	IRF2	IRF2BP1	IRF4	IRF5	ITGA2B	ITGAV	ITIH4

596	ITPR2	ITSN1	JARID2	JOSD1	JUP	KARS	KAT7
603	KBTBD11	KCNC3	KCNG1	KCNMB1	KCNQ3	KCTD14	KCTD20
610	KDELC2	KDM3A	KIAA0040	KIF15	KIF3B	KIN	KIR2DL1
617	KLF1	KLF6	KLF7	KLHDC10	KLHL14	KLHL18	KLHL21
624	KLK13	KLRC1	KLRC2	KLRC3	KMT2D	KRT72	KXD1
631	KYNU	LACTB	LAIR1	LAMP5	LDHB	LEAP2	LETMD1
638	LGALS2	LGALS9	LGALS9B	LGALS9C	LIG4	LILRA1	LILRA3
645	LILRB2	LILRB4	LIMK1	LIMS1	LIMS3	LIN9	LINC00467
652	LLGL1	LMNB1	LNPEP	LPAR1	LPAR6	LPGAT1	LPIN2
659	LRPPRC	LRRC3	LRRFIP1	LTA4H	LY6G5B	MACF1	MAK
666	MALT1	MAMLD1	MAN1A1	MAN1A2	MAN2A2	MAOB	MAP1B
673	MAP2K6	MAP3K2	MAP3K3	MAP3K4	MAP3K5	MAP3K9	MAP4K2
680	MAPK14	MAPKAP1	MAPRE2	MARCKS	MARCO	MAST1	MAST2
687	MAST3	MBD1	MBD4	MBNL2	MBNL3	MCAT	MCL1
694	MCM2	MCM5	MDK	MDM1	MDM2	MED12	MED22
701	MEF2D	MEFV	MFNG	MFSD5	MIA3	MIEF1	MIR600HG
708	MLKL	MLXIP	MMP12	MNDA	MNT	MOAP1	MOB3C
715	MOBP	MPHOSPH9	MPPE1	MRPL40	MRPS27	MSL3	MSMO1
722	MSR1	MT1G	MT1X	MTDH	MTHFD1	MTMR6	MTSS1
729	MVB12A	MVP	MX1	MX2	MXD1	MYBPC3	MYC
736	MYCL	MYCN	MYOF	N4BP1	N4BP3	NADK	NAGK
743	NAP1L5	NAPSA	NAPSB	NARS	NASP	NAV3	NBEAL1
750	NBN	NBR1	NCAPD2	NCOA7	NDC80	NDUFS1	NEDD4
757	NET1	NETO2	NEXN	NFAT5	NFATC3	NFE2	NFKBIL1
764	NFRKB	NINJ1	NINJ2	NISCH	NME5	NME8	NONO
771	NOP56	NPAT	NPC1	NPC2	NPM1	NPR1	NPRL2
778	NR1D2	NR4A1	NRG1	NRIP1	NRXN3	NSUN3	NSUN7
785	NT5C2	NTNG2	NTRK3	NUMA1	NUP210	NUP214	NUP98
792	NUSAP1	NXF1	NXT2	OAS3	OASL	OAT	ODC1
799	ODF3B	OGFR	OGG1	OLFM1	OLIG2	OPTN	OSBPL1A
806	OSGEPL1	OTOF	OTUD1	OXT	P2RX1	P2RX7	P2RY10
813	P2RY13	P2RY14	PABPC1	PABPC4	PALLD	PALM2-AKAP2	PANK3
820	PAPD7	PAPSS2	PARD6A	PARK7	PARP11	PARP14	PARP2
827	PATL1	PAX3	PCID2	PCNA	PCOLCE2	PDE1C	PDGFRA
834	PDGFRL	PDK1	PDSS1	PDXK	PEG3	PEX7	PGAP1
841	PGM1	PHACTR2	PHACTR4	PHF20	PHLDA2	PI4K2A	PI4K2B
848	PIK3AP1	PIK3CB	PIK3CD	PIK3CG	PIK3IP1	PIM1	PIN4
855	PINK1-AS	PITPNC1	PIWIL4	PLA2G15	PLAA	PLAGL1	PLD2
862	PLEKHA1	PLEKHB2	PLEKHN1	PLGRKT	PLK2	PLN	PLSCR4
869	PLXNA2	PLXNB2	PMAIP1	PMEPA1	PMS1	PNISR	PNPT1
876	POLG2	POLI	POLK	POLR2C	POLR2J	POLR2J2	POLR2J3
883	POLR2J4	POP7	PPBP	PPDPF	PPFIBP2	PPIA	PPM1F
890	PPM1H	PPM1K	PPP1CC	PPP2R2A	PRADC1	PRAM1	PRC1

897	PRDM2	PRF1	PRIM1	PRIMPOL	PRKACB	PRKAG2	PRKAR2A
904	PRKCE	PRKD2	PRKDC	PRPF8	PRPS2	PRPSAP1	PSD3
911	PSIP1	PSMA2	PSMA4	PSMC3IP	PTGES	PTGFRN	PTK2
918	PTP4A1	PTP4A2	PTPN13	PTPN22	PTPN4	PTPRCAP	PTPRK
925	PTPRO	PURA	PYGL	QPCT	QSOX2	RAB11FIP2	RAB22A
932	RAB27A	RAB32	RAB37	RAB38	RAB40A	RAB40AL	RAB40C
939	RAB5B	RAB8A	RAB8B	RABGGTB	RAD9A	RAI14	RALB
946	RALBP1	RANGAP1	RAP2B	RARB	RASA4	RASA4CP	RASAL1
953	RASGRP3	RASSF1	RASSF2	RB1	RBCK1	RBL2	RBM11
960	RBM3	RBM43	RBM4B	RBM8A	RBMS1	RBMS3	RBMX
967	RBMY1HP	RBMY1J	RECK	REL	RELA	REM1	RFT1
974	RGL1	RGL2	RGP1	RGS1	RGS2	RHBDD2	RHOB
981	RHOF	RICTOR	RIF1	RILPL2	RIN2	RIN3	RIPK1
988	RIPK2	RLIM	RMND5A	RNASE1	RNASET2	RNF113A	RNF144A
995	RNF170	RNF213	RNF41	RNF44	RNMT	RNU7-1	RP2
1002	RPL11	RPL13	RPL13A	RPL19	RPL22	RPL23	RPL3
1009	RPL30	RPL34	RPL4	RPL5	RPLP0P6	RPP40	RPRD2
1016	RPS27L	RPS4Y1	RPS6KA2	RPS8	RRM2	RRP8	RSBN1
1023	RSL1D1	RTCB	RTN3	RUNX3	RWDD1	S100A4	S100A8
1030	S100A9	SAG	SAMD4A	SAMHD1	SAR1B	SART1	SART3
1037	SAT1	SBNO2	SCAPER	SCARB2	SCIMP	SCRN3	SEC24D
1044	SECTM1	SELL	SEMA3C	SEMA4F	SEMA7A	SEPHS1	SEPHS2
1051	SEPT2	SERPINF2	SERTAD2	SETDB2	SFT2D2	SGMS2	SH2B2
1058	SH2D3C	SH3BP5	SHISA5	SHMT2	SIGLEC1	SIGLEC7	SLA
1065	SLAMF8	SLC10A1	SLC11A1	SLC16A1	SLC16A5	SLC18B1	SLC19A1
1072	SLC22A13	SLC22A5	SLC24A4	SLC25A20	SLC25A24	SLC25A25	SLC25A36
1079	SLC25A39	SLC30A1	SLC31A2	SLC35D2	SLC36A1	SLC38A5	SLC39A8
1086	SLC43A2	SLC48A1	SLC4A7	SLC5A3	SLC7A1	SLC7A6	SLC8A1
1093	SLC9B2	SLFN12	SLFN13	SLFN5	SMAD3	SMAD6	SMCHD1
1100	SMG7	SMO	SNAP23	SNAPC4	SNHG1	SNRNP200	SNRPN
1107	SNX29P2	SNX32	SOCS2	SOCS7	SOS1	SOS2	SOWAHC
1114	SOX4	SP100	SP110	SP140L	SPAG7	SPATS2L	SPCS3
1121	SPEN	SPI1	SPOCK2	SPTA1	SPTBN1	SPTLC2	SPTSSA
1128	SQLE	SREBF2	SRGAP2	SRGAP2B	SRGAP2C	SRPK1	SRPX2
1135	SSH1	SSSCA1	SSTR2	ST3GAL1	ST3GAL5	ST3GAL6	STAG3L3
1142	STAM	STAP1	STARD13	STARD4	STARD7	STK17A	STK3
1149	STK38L	STOM	STOML1	STRBP	STX17	STX3	SUN2
1156	SUPT7L	SUPV3L1	SV2A	SWAP70	SYNE1	SYNGR3	TACSTD2
1163	TAF10	TAOK2	TAP2	TARBP1	TAS2R5	TBCEL	TBL1XR1
1170	TBR1	TBX6	TBXAS1	TCF7L2	TCFL5	TCP1	TDRD7
1177	TFAP4	TFEB	TFEC	TFF2	TFF3	TGFA	TGFBR1
1184	TGFBR2	TGM1	TGM5	TGOLN2	THAP11	THBS1	THPO
1191	THUMPD3	TIGD5	TIMP3	TIPARP	TIPIN	TKT	TKTL1
1198	TLK2	TLR3	TLR4	TLR6	TMBIM1	TMEM106A	TMEM110

1205	TMEM110-MUSTN1	TMEM120B	TMEM123	TMEM126B	TMEM135	TMEM140	TMEM141
1212	TMEM186	TMEM19	TMEM223	TMEM242	TMEM243	TMEM30A	TMEM55A
1219	TMEM62	TMEM97	TMOD2	TMX4	TNFRSF1B	TNFSF18	TNFSF9
1226	TNK2	TOB1	TOMM20	TOP1MT	TOP2B	TOX	TP53
1233	TPCN1	TPP1	TPST2	TRA2A	TRABD	TRAF1	TRAF3IP3
1240	TRAIP	TRANK1	TREM1	TRIAP1	TRIB2	TRIL	TRIM14
1247	TRIM25	TRIM26	TRIM34	TRIM38	TRIM5	TRIM56	TRIM6-TRIM34
1254	TRIM65	TRIP10	TRIP4	TRIP6	TRPS1	TSPAN13	TSPO
1261	TSPYL5	TTC21A	TTC28	TTC39B	TTC7B	TTLL4	TTN
1268	TUFT1	TXNDC12	TYMS	UBA7	UBAC1	UBASH3B	UBE2G1
1275	UBE2S	UBE2W	UBE4B	UBL4A	UBQLNL	UBXN2A	UBXN7
1282	UCP3	UEVLD	UNC93B1	UQCC1	USP15	USP18	USP22
1289	USP25	USP4	USP41	USP6	USP6NL	UTRN	VASH1
1296	VCAN	VCY1B	VEGFA	VENTX	VPREB3	VPS9D1	VRK2
1303	VTN	WBP1L	WDFY1	WDR25	WDR74	WIPF1	WISP1
1310	WNK1	XIAP	YPEL3	YY1	ZAP70	ZBED1	ZBTB18
1317	ZBTB48	ZC3HAV1	ZCCHC2	ZFAND2A	ZFP36	ZFP36L2	ZFP69B
1324	ZFYVE26	ZNF101	ZNF107	ZNF146	ZNF211	ZNF23	ZNF248
1331	ZNF264	ZNF267	ZNF280B	ZNF318	ZNF324	ZNF350	ZNF443
1338	ZNF446	ZNF512	ZNF516	ZNF543	ZNF552	ZNF618	ZNF626
1345	ZNF652	ZNF688	ZNF702P	ZNF75D	ZNF780A	ZNF780B	ZNF814
1352	ZNF85	ZNFX1	ZWINT				

Supplementary Table 3 IFN type II signaling genes present in MDS according to Interferome database classification

1	AADAT	AATK	ABCC2	ABHD6	ABLIM1	ACAT2	ACBD5
8	ACHE	ACOT1	ACOT7	ACSS2	ACTG1	ACTN1	ACVR1B
15	ADAM11	ADAM9	ADCY1	ADO	ADORA2A	AFAP1L1	AK2
22	AK8	AKR1B1	AKR7A2P1	ALCAM	ALDH1A1	ALDH2	ALOX5
29	ALPI	AMDHD1	AMELX	AMELY	ANKRD11	ANLN	ANO9
36	ANTXR1	ANXA2P3	AP1B1	AP5B1	APBB1IP	APCS	APOL2
43	APOO	AR	ARID5B	ARMC9	ARNTL2	ARSB	ASAP1
50	ASPHD1	ASPHD2	ASPM	ATG4C	ATP1A1	ATP5F1	ATP6V0A1
57	ATP6V0D2	ATP6V1H	ATP8B4	AURKA	AVPI1	B3GNT5	B3GNT8
64	BAG3	BAMBI	BASP1	BCAP31	BCAR3	BCAT1	BCL11A
71	BHLHE22	BHLHE40	BIN1	BMP6	BOLA3	BRSK1	BTBD11
78	BTN3A1	C10orf95	C11orf45	C11orf96	C18orf8	C19orf12	C1orf112
85	C1orf61	C1QB	C1QC	C1QTNF1	C1R	C1RL	C1S
92	C2	C3orf58	C4orf3	C4orf32	C6orf223	C6orf47	CA2
99	CA9	CABLES1	CACNA2D3	CADM1	CALCA	CALCOCO2	CALM3
106	CALML4	CAMKK1	CAMP	CASK	CASP5	CBX4	CCDC115
113	CCL18	CCL3L1	CCL3L3	CCL5	CCM2L	CCNA2	CCNB1
120	CCNB2	CCND1	CCND2	CCNE2	CCNO	CD109	CD1A
127	CD209	CD274	CD276	CD300LB	CD47	CD58	CD72
134	CDC20	CDC45	CDCA5	CDK18	CDKN3	CDR2	CDYL2
141	CEBPD	CELF1	CENPU	CENPW	CEP135	CEP19	CEP55
148	CES1	CETP	CFP	CHD2	CHDH	CHI3L2	CHRNA6
155	CHST7	CIITA	CIT	CKAP2	CKAP2L	CLDN12	CLDND1
162	CLEC12A	CLEC7A	CLIP4	CLPS	CLPTM1	CLSPN	CMC4
169	CMTM3	CNIH3	CNIH4	CNRIP1	CNTFR	COCH	COL4A1
176	COL8A2	COMMD6	COQ10B	CORO2A	CR2	CRABP2	CRYBB1
183	CSF3	CSPG4	CSRNP2	CTDSPL	CTNNBIP1	CTNND2	CTSK
190	CTSO	CTSS	CTTN	CXADR	CXorf57	CYB5R3	CYFIP1
197	CYSLTR1	CYTH4	CYTIP	DAPP1	DAXX	DBN1	DDA1
204	DDIT4	DENND4C	DEXI	DGKD	DHRS11	DHRS13	DHRS9
211	DIRAS1	DISP1	DLGAP5	DNAJB6	DNAJC9	DNASE2B	DNM3
218	DNMT3A	DOCK2	DOCK3	DOPEY2	DPEP2	DPYD	DPYSL4
225	DRAM1	DUSP14	DUSP2	DUSP6	E2F2	EAF2	EBI3
232	ECM1	EDA	EDEM1	EDN1	EEPD1	EFR3B	EMB
239	EMILIN2	EPAS1	EPB41L3	EPN2	ERICH1	ETV5	EVL
246	F13A1	FABP3	FAH	FAIM	FAM109A	FAM117B	FAM124A
253	FAM129A	FAM129B	FAM135A	FAM20A	FAM20C	FAM213A	FAM81A
260	FANCG	FAXDC2	FBXO15	FCAMR	FCER1G	FEZ2	FGD3
267	FGD5	FGF13	FGFRL1	FGGY	FHL1	FKBP5	FLI1
274	FLNB	FLOT2	FMNL2	FNDC3A	FOSL2	FOXN2	FOXN3
281	FOXRED2	FRMD4A	FSCN1	G0S2	GAA	GABBR1	GALNT12
288	GAPT	GAS2L3	GAS6	GATA1	GCLC	GCSHP5	GEM
295	GFOD1	GGTA1P	GIMAP1	GIMAP2	GIMAP4	GIMAP5	GIMAP7
302	GIMAP8	GINM1	GJC1	GLCCI1	GLDN	GLO1	GLRX5

309	GLT1D1	GNB1L	GNE	GNG2	GOLIM4	GPER1	GPR132
316	GPR162	GPR31	GPR34	GPR84	GPRC5B	GPRIN3	GRAMD1A
323	GRINA	GSTT2	GSTT2B	GUCY1B3	GUCY2F	GVINP1	GYS1
330	H2AFY2	HBEGF	HCK	HCST	HDAC4	HDAC6	HEATR6
337	HERPUD1	HGSNAT	HILPDA	HIPK1	HIST1H2AM	HIST1H2BG	HIST1H2BI
344	HIST1H2BJ	HIST1H2BL	HIST1H2BM	HIST1H3A	HIST1H3C	HIST1H3J	HIST1H4B
351	HIST2H2AB	HIVEP3	HLA-A	HLA-DMA	HLA-DMB	HLA-DOA	HLA-DPB1
358	HLA-DRA	HLTF	HLX	HMG20B	HMGCS1	HNRNPDL	HOMER3
365	HOXD3	HPCAL1	HS3ST2	HS3ST3B1	HSD11B1	HSD17B7	HSPA6
372	HTR7	HYI	ICAM3	IDO2	IFT172	IGF1	IL10RB
379	IL12RB1	IL13RA1	IL18BP	IL1R2	IL32	IL4R	ILDR1
386	INPP4A	IQCE	IRS2	ISLR2	ITGB1BP1	ITGB7	ITPKC
393	JAK3	JAKMIP2	KAT2A	KCNA1	KCNAB2	KCNH1	KCNJ10
400	KCNK13	KCNMA1	KCNQ1	KCTD17	KIAA1147	KIF13B	KLF4
407	KLF9	KLK3	KLRG1	KPNA2	KRBA1	KREMEN1	LAMA3
414	LASP1	LAT	LDB3	LFNG	LGALSL	LIG1	LILRB5
421	LIMA1	LIMK2	LINC00324	LITAF	LMNA	LONP1	LPAR2
428	LRG1	LRP12	LRRC38	LRRCC1	LRRFIP2	LRRK2	LSM3
435	LXN	LYPD5	LYRM9	MAD2L1	MAD2L1BP	MAFB	MAFF
442	MANBAL	MAP2K3	MAP3K8	MAP7	MAPK13	MARCH1	MARCH3
449	MATK	ME1	ME3	MELK	METTL1	METTL7A	MFSD12
456	MFSD3	MGLL	MGST1	MIA	MICAL1	MICAL2	MICALL2
463	MIR22HG	MIS18BP1	MKL2	MLLT6	MLPH	MMD	MME
470	MMP10	MMP25	MORN2	MOXD1	MPEG1	MPP1	MPZL2
477	MR1	MRC1	MREG	MRPL12	MRPL49	MSANTD3	MSC
484	MSRB2	MTCP1	MTFP1	MTHFS	MTUS1	MUCL1	MVB12B
491	MYH11	MYL5	MYLIP	MYO1D	MYO1G	MYOC	MZT2A
498	NAA16	NAF1	NCAPG	NCF4	NDP	NDRG1	NDUFAF2
505	NEDD4L	NEIL1	NEK6	NF1	NFIA	NFKB1	NFKB2
512	NFKBIA	NFKBIB	NME4	NNMT	NOVA1	NPM3	NR2F6
519	NRIP3	NUMB	NUP85	OLFML3	OPRL1	OSBPL9	OSMR
526	OSTC	P2RY8	PACS2	PAG1	PAPLN	PAPOLG	PARP3
533	PASK	PAWR	PBDC1	PBX3	PCBP2	PDE3B	PDE9A
540	PDGFB	PDIA4	PDIA6	PEA15	PEAK1	PEMT	PFKFB3
547	PFN1P6	PGD	PGS1	PHIP	PHLDA3	PIGQ	PIK3C2B
554	PIM3	PLA1A	PLAC8	PLAGL2	PLAT	PLAUR	PLCXD1
561	PLEKHG2	PLEKHG3	PLTP	PM20D2	PMFBP1	PNKD	PNMT
568	POMT2	PPA1	PPIAL4C	PPIAL4D	PPIAL4G	PPIC	PPP1R14B
575	PPP2R5B	PPTC7	PRELID1	PRKAR2B	PROCR	PRR13	PRUNE2
582	PSEN2	PSMA5	PSMA6	PSMB10	PSTPIP2	PTEN	PTGER2
589	PTGIR	PTGR1	PTPN7	PTPRE	QPRT	RAB12	RAB21
596	RAB30	RAB3IL1	RAB42	RAP2A	RAP2C	RAPGEF1	RARA
603	RARRES1	RASL10A	RBBP9	RBFA	RBP1	RBP7	RCN3
610	RDH10	RDH11	RELB	RFTN1	RFX1	RFX2	RFX5
617	RGS10	RGS12	RHBDF1	RHOBTB2	RHOBTB3	RILPL1	RIMBP3
624	RIMBP3C	RIMS3	RND1	RNF114	RNF141	RNF149	RNF24

631	RPGRIP1	RRAD	RTN4R	SACS	SAMD8	SAR1A	SAV1
638	SCP2	SDC1	SDC3	SEMA4B	SEMA4D	SEPT11	SEPT8
645	SERPINB3	SERPINB4	SERPINF1	SESN2	SGPP2	SH2D4A	SH3BGRL3
652	SH3PXD2A	SH3PXD2B	SH3RF1	SH3RF3	SHC3	SIX1	SLAMF6
659	SLAMF9	SLC25A10	SLC25A37	SLC26A11	SLC27A3	SLC28A3	SLC35E3
666	SLC39A10	SLC39A14	SLC3A1	SLC40A1	SLC43A3	SLC45A3	SLC46A1
673	SLC50A1	SLC5A4	SLC6A6	SLC7A8	SLC8A3	SLC9A1	SLCO4A1
680	SLCO5A1	SLPI	SMAD7	SMARCA1	SMARCD3	SMCO4	SNX10
687	SNX21	SNX27	SOCS3	SPATA13	SPECC1L-ADORA2A	SPEG	SPHK1
694	SPHK2	SPN	SPRED1	SPRY2	SRC	SRSF12	SRXN1
701	SSH2	ST20-MTHFS	ST6GALNAC6	STAB1	STARD8	STAT4	STEAP3
708	STIP1	STK32C	STMN1	STX10	SV2B	SYNE3	TAF13
715	TAF1A	TANC1	TAPBPL	TBC1D1	TBC1D12	TBC1D2B	TBC1D4
722	TBC1D9	TBCD	TCF4	TCTEX1D2	TDRD9	TEAD4	TEX2
729	TEX33	TFB2M	TGFB3	THADA	THBD	TIAM2	TIMP1
736	TJP2	TLE4	TLR8	TM4SF19	TMC8	TMCC2	TMED3
743	TMED7-TICAM2	TMEM106C	TMEM14C	TMEM165	TMEM170A	TMEM217	TMEM43
750	TMEM50A	TMEM80	TMEM86A	TMEM8B	TMSB15A	TNFAIP2	TNFRSF12A
757	TNFRSF21	TNIP2	TNIP3	TNKS	TOP2A	TP53INP2	TPM2
764	TPM3	TRADD	TRAF3IP2	TREM2	TRIP13	TRPC4AP	TRPV2
771	TSC22D1	TSFM	TSPAN14	TTC8	TTC9C	TUBB4B	TUBB6
778	TULP4	TUT1	TXNIP	TXNRD1	TXNRD3	UBE2C	UBE2E3
785	UBFD1	UBTD1	UBXN11	UCHL3	UFSP2	UGCG	UHRF1
792	UNG	UNKL	UTS2	VASN	VAT1	VCP	VCPIP1
799	VPS26B	VWF	WFDC2	WTAP	XPNPEP1	YBX3	ZBED3
806	ZBED6CL	ZBTB17	ZBTB7A	ZMIZ2	ZMYM6	ZNF219	ZNF366
813	ZNF573	ZNF79	ZNRF3	ZYX			

Supplementary Table 4 INF I and II inducible genes identified by Interferome v2.0 database

1	ABCC5	ABCG1	ACOX2	ACP2	ADAR	ADK	ADM
8	ADORA2B	AHNAK	AIM2	AK4	ALDOC	AMPD3	ANKRD22
15	APOBEC3G	APOL1	APOL3	APOL6	AQP9	ARAP3	ASCL2
22	ATF3	ATOX1	ATP10A	AXL	B4GALT5	BAK1	BATF
29	BATF2	BAZ1A	BCL2L13	BIRC3	BLOC1S5-TXNDC5	BLZF1	BRCA2
36	BRMS1L	BTG3	BTN3A2	BTN3A3	C15orf48	C1QA	C20orf27
43	C5orf56	CAMK1G	CASP3	CCL19	CCL23	CCL4	CCL7
50	CCL8	CCR2	CCR7	CD163L1	CD180	CD1D	CD302
57	CD36	CD38	CD40	CD44	CD52	CD74	CD83
64	CD9	CD93	CDA	CDK1	CDKN1A	CENPM	CFB
71	CFH	CFLAR	CHST2	CLEC5A	CMTM7	CNP	COLEC12
78	CREBRF	CREM	CSF2RA	CSF2RB	CSTB	CTSC	CTSL
85	CXCL11	CXCL13	CXCL9	CYB561A3	CYP27B1	CYSTM1	CYTL1
92	DAB2	DBI	DBP	DCSTAMP	DDX60	DENND5A	DHX58
99	DNAJA1	DNASE2	DSE	DUSP1	DUSP10	DYNLT1	DYRK2
106	EIF2AK2	EIF4EBP2	EMP1	ENPP2	EPSTI1	ETV7	EZR
113	FABP4	FAM105A	FAM26F	FAM53B	FAM60A	FAS	FBP1
120	FBXO6	FCGR1A	FCGR1B	FCN1	FGFR2	FGL2	FGR
127	FPR2	FZD2	FZD5	GATM	GBP1P1	GBP2	GBP4
134	GBP5	GCH1	GCNT1	GCSH	GDF15	GIMAP6	GLS
141	GM2A	GPC4	GPD1L	GPR155	GRIN3A	GSDMD	GSN
148	GTF2B	GTPBP1	GTPBP2	GUCY1A3	HAPLN3	HCP5	HHEX
155	HIVEP2	HK3	HLA-C	HLA-F	HMMR	HPN	HTRA1
162	HVCN1	IDO1	IFI35	IFI44L	IFI6	IFIT1	IFIT3
169	IFIT5	IFITM1	IFITM2	IFITM3	IL16	IL1RN	IL2RA
176	IL3RA	IL7R	INF2	IQCK	IRAK3	IRF1	IRF7
183	IRF9	ISG15	ISG20	ITGAE	ITGB5	ITPKB	IVNS1ABP
190	JADE2	JAK2	KCNJ2	KCTD12	KIAA0513	KLHDC7B	KRT85
197	LAG3	LAMP3	LAP3	LCP2	LDLR	LDLRAD4	LEPROTL1
204	LGALS3BP	LGMN	LILRA5	LMO4	LPCAT1	LPL	LY6E
211	LY9	LYN	LYRM1	LYSMD2	MAD2L2	MAF	MARCKSL1
218	MASTL	MB21D1	MCM6	MCOLN2	MEF2C	MEGF9	MERTK
225	MICB	MITF	MLEC	MLXIPL	MMP7	MMP9	MOV10
232	MS4A4A	MS4A6A	MT1E	MT1F	MT1M	MTHFD2	MUC1
239	MXD4	MYD88	NABP1	NACC2	NAMPT	NAP1L1	NAPA
246	NCF1C	NCR1	NEURL3	NFE2L3	NFIL3	NMI	NOD1
253	NOD2	NOTCH1	NQO1	NR1H3	NR4A3	NRGN	NRP1
260	NT5C3A	NUB1	NUPR1	OAS1	OAS2	OIP5	P2RY6
267	PAM	PANX1	PAPSS1	PARP10	PARP12	PARP9	PC
274	PCED1B	PCGF5	PDE4B	PDLIM7	PDPN	PFKFB4	PFKP
281	PHF11	PHLDA1	PID1	PILRA	PLA2G16	PLA2G7	PLIN2
288	PLSCR1	PLXDC2	PML	PMP22	PNP	PNRC1	POLB
295	PON2	PPARG	PPFIBP1	PPIF	PRDM1	PRKCA	PRNP
302	PSMB8	PSMB9	PSME1	PSME2	PTGS1	PTPN1	PTPN2

309	RAC2	RAD51AP1	RAPH1	RARRES3	RASGRP1	RBBP6	RCSD1
316	RHOH	RNF144B	RNF19B	RPL15	RPS6KA1	RSAD2	RTN1
323	RTP4	RXRA	SAMD9	SAMD9L	SAMM50	SAMSN1	SAP30
330	SASH1	SCG5	SCN1B	SCPEP1	SDC2	SECISBP2L	SEL1L3
337	SEMA4A	SEPT4	SERPINA1	SERPINB1	SERPINB9	SERPING1	SLAMF7
344	SLC16A10	SLC25A28	SLC2A1	SLC41A2	SLC9A7	SMPD3	SOCS1
351	SOD2	SORL1	SP140	SPARC	SPOCD1	SPP1	SPRY1
358	SPSB1	ST6GALNAC4	ST8SIA4	STAMBPL1	STAT1	STAT2	STAT3
365	STRN	STX11	TAGAP	TANK	TAP1	TBC1D10C	TBC1D22B
372	TBK1	TCN2	TCP11L1	TDRD3	TG	TGM2	TK1
379	TLR5	TLR7	TMEM158	TNF	TNFAIP3	TNFAIP6	TNFRSF11A
386	TNFSF10	TNFSF13B	TNFSF14	TNIP1	TNNI2	TOP1	TOR1B
393	TPST1	TRAFD1	TRIM21	TRIM22	TRIM69	TRPM2	TSC22D3
400	TUBGCP4	TYMP	UBE2D1	UBE2L6	UCP2	ULK2	UPP1
407	UTS2R	VAMP5	VPS41	VSIG10L	VSIG4	VWA5A	WARS
414	WWP1	XAF1	XRN1	YEATS2	YPEL2	ZBP1	ZBTB16
421	ZCCHC14	ZHX3	ZMIZ1	na	na	na	na

Supplementary Table 5 Module enrichment in the genes sorted by weights in PC1, PC2 and PC8 of gene expression set of the MDS

ID	Title	AUC	adj.P.Val
	Enrichment in PC1		
LI.M7.0	enriched in T cells (I)	0.935812	1.41E-29
LI.M7.1	T cell activation (I)	0.909565	1.42E-20
LI.M7.2	enriched in NK cells (I)	0.829489	3.07E-12
LI.M7.4	T cell activation (III)	0.972746	5.92E-10
LI.M18	T cell differentiation via ITK and PKC	0.975338	8.45E-10
LI.S0	T cell surface signature	0.842331	1.61E-08
LI.M7.3	T cell activation (II)	0.834783	1.79E-08
LI.M106.0	nuclear pore complex	0.951424	3.82E-08
LI.M14	T cell differentiation	0.942652	1.02E-07
LI.M5.1	T cell activation and signaling	0.801434	1.23E-07
LI.M212	purine nucleotide biosynthesis	0.93928	1.63E-07
LI.M245	translation initiation factor 3 complex	0.9521	2.73E-07
LI.M61.2	enriched in NK cells (receptor activation)	0.8859	1.04E-06
LI.M157	enriched in NK cells (III)	0.925799	1.57E-06
LI.M19	T cell differentiation (Th2)	0.865283	4.87E-06
LI.M117	cell adhesion (GO)	0.702055	5.69E-06
LI.M181	nucleotide metabolism	0.93683	5.90E-06
LI.M61.0	enriched in NK cells (II)	0.850627	7.51E-06
LI.M235	mitochondrial cluster	0.900809	1.14E-05
LI.M143	nuclear pore, transport; mRNA splicing, processing	0.936675	2.27E-05
LI.M47.0	enriched in B cells (I)	0.809911	5.00E-05
LI.M204.0	chaperonin mediated protein folding (I)	0.85285	5.78E-05
LI.M65	IL2, IL7, TCR network	0.803117	6.17E-05
LI.M169	mitosis (TF motif CCAATNNSNNNGCG)	0.870433	0.000156
LI.M106.1	nuclear pore complex (mitosis)	0.883762	0.000185
LI.M223	enriched in T cells (II)	0.926972	0.000309
LI.M4.8	cell division - E2F transcription network	0.828515	0.000311
LI.M103	cell cycle (III)	0.729376	0.000446
LI.M4.4	mitotic cell cycle - DNA replication	0.788318	0.000766
LI.S7	CD4 T cell surface signature Th2-stimulated	0.786311	0.001363
LI.M126	double positive thymocytes	0.788578	0.001363
LI.M47.1	enriched in B cells (II)	0.766986	0.001531
LI.M182	enriched in DNA interacting proteins	0.792062	0.001531
LI.M10.0	E2F1 targets (Q3)	0.709764	0.002685
LI.M204.1	chaperonin mediated protein folding (II)	0.816658	0.004354
LI.M179	enriched for TF motif PAX3	0.880389	0.005378
LI.M22.0	mismatch repair (I)	0.753984	0.005429
LI.M130	enriched in G-protein coupled receptors	0.833819	0.005429
LI.M5.0	regulation of antigen presentation and immune response	0.498417	0.006667
LI.M4.1	cell cycle (I)	0.670012	0.007241
LI.M45	leukocyte activation and migration	0.694664	0.007241

LI.M69	enriched in B cells (VI)	0.803069	0.007529
LI.M12	CD28 costimulation	0.770624	0.008251
LI.M60	lymphocyte generic cluster	0.75206	0.009
LI.M156.0	plasma cells & B cells, immunoglobulins	0.808926	0.009
LI.M47.4	enriched in B cells (V)	0.756107	0.009309
LI.M10.1	E2F1 targets (Q4)	0.780227	0.009863
LI.M76	DNA repair	0.736904	0.010472
LI.M230	cell cycle, mitotic phase	0.741561	0.012543
LI.M44	T cell signaling and costimulation	0.668879	0.014104
LI.M22.1	mismatch repair (II)	0.778028	0.019561
LI.M4.11	mitotic cell cycle in stimulated CD4 T cells	0.777793	0.01989
LI.M156.1	plasma cells, immunoglobulins	0.860013	0.023462
LI.M250	spliceosome	0.762262	0.034335
LI.M37.3	cell division	0.837723	0.034443
	Enrichment in PC2		
LI.M11.0	enriched in monocytes (II)	0.806343	2.06E-19
LI.M37.0	immune activation - generic cluster	0.679038	1.07E-12
LI.M4.0	cell cycle and transcription	0.712427	1.53E-11
LI.M237	golgi membrane (II)	0.962035	1.53E-11
LI.S4	Monocyte surface signature	0.793659	6.04E-09
LI.M213	regulation of transcription, transcription factors	0.895308	2.52E-08
LI.M101	phosphatidylinositol signaling system	0.944165	1.06E-07
LI.M129	inositol phosphate metabolism	0.931895	2.11E-07
LI.M144	cell cycle, ATP binding	0.950882	2.53E-07
LI.M118.0	enriched in monocytes (IV)	0.78088	2.59E-07
LI.M16	TLR and inflammatory signaling	0.842178	9.29E-07
LI.M37.1	enriched in neutrophils (I)	0.815243	1.03E-06
LI.M147	intracellular transport	0.873479	1.23E-05
LI.M73	enriched in monocytes (III)	0.893751	0.001746
LI.M5.0	regulation of antigen presentation and immune response	0.634725	0.001862
LI.M64	enriched in activated dendritic cells/monocytes	0.852277	0.002614
LI.M169	mitosis (TF motif CCAATNNSNNNGCG)	0.823067	0.00416
LI.M165	enriched in activated dendritic cells (II)	0.744301	0.00416
LI.M113	golgi membrane (I)	0.88044	0.005106
LI.M11.1	blood coagulation	0.801502	0.005845
LI.M138	enriched for ubiquitination	0.87511	0.006183
LI.M22.0	mismatch repair (I)	0.767709	0.006303
LI.M4.3	myeloid cell enriched receptors and transporters	0.768503	0.009813
LI.M179	enriched for TF motif PAX3	0.860471	0.012319
LI.M3	regulation of signal transduction	0.679308	0.014948
LI.M4.13	cell junction (GO)	0.865614	0.015924
LI.M132	recruitment of neutrophils	0.871681	0.015924
LI.M25	TLR8-BAFF network	0.831208	0.019401
LI.M53	inflammasome receptors and signaling	0.787681	0.019401
LI.M81	enriched in myeloid cells and monocytes	0.620112	0.019401

LI.M226	proteasome	0.805515	0.027548
LI.M163	enriched in neutrophils (II)	0.811818	0.027548
LI.M230	cell cycle, mitotic phase	0.7676	0.029278
LI.M40	complement and other receptors in DCs	0.75148	0.029278
LI.M191	transmembrane transport (II)	0.658033	0.030142
LI.M114.1	glycerophospholipid metabolism	0.784245	0.032203
LI.S11	Activated (LPS) dendritic cell surface signature	0.676177	0.032203
LI.M86.0	chemokines and inflammatory molecules in myeloid cells	0.713533	0.034613
LI.M118.1	enriched in monocytes (surface)	0.776815	0.03509
LI.M56	suppression of MAPK signaling	0.813592	0.04503
LI.M127	type I IFN response	0.825481	0.048774
	Enrichment in PC8		
LI.M4.0	cell cycle and transcription	0.646351	2.05E-16
LI.M4.1	cell cycle (I)	0.782294	2.05E-16
LI.M127	type I IFN response	0.996512	1.72E-15
LI.M75	antiviral IFN signature	0.802929	2.42E-15
LI.M150	innate antiviral response	0.910664	3.28E-14
LI.M165	enriched in activated dendritic cells (II)	0.692141	2.21E-13
LI.M111.1	viral sensing & immunity; IRF2 targets network (II)	0.992409	8.48E-13
LI.M67	activated dendritic cells	0.961194	1.26E-11
LI.M146	MHC-TLR7-TLR8 cluster	0.908728	5.35E-09
LI.M111.0	viral sensing & immunity; IRF2 targets network (I)	0.66489	5.05E-08
LI.M7.2	enriched in NK cells (I)	0.76298	7.27E-08
LI.M4.5	mitotic cell cycle in stimulated CD4 T cells	0.808153	2.16E-07
LI.M4.4	mitotic cell cycle - DNA replication	0.837511	1.58E-06
LI.M103	cell cycle (III)	0.745543	1.69E-06
LI.M13	innate activation by cytosolic DNA sensing	0.8076	3.35E-06
LI.S1	NK cell surface signature	0.687714	8.49E-06
LI.M68	RIG-1 like receptor signaling	0.764642	3.15E-05
LI.M10.0	E2F1 targets (Q3)	0.758543	8.27E-05
LI.M4.7	mitotic cell cycle	0.873012	0.000104
LI.M4.2	PLK1 signaling events	0.78707	0.000105
LI.M46	cell division stimulated CD4+ T cells	0.759223	0.000105
LI.M7.3	T cell activation (II)	0.735883	0.000148
LI.M200	antigen processing and presentation	0.94937	0.000414
LI.M22.0	mismatch repair (I)	0.755343	0.00054
LI.M4.8	cell division - E2F transcription network	0.789056	0.000608
LI.M4.6	cell division in stimulated CD4 T cells	0.754704	0.000908
LI.S11	Activated (LPS) dendritic cell surface signature	0.634705	0.000922
LI.M119	enriched in activated dendritic cells (I)	0.820204	0.000939
LI.M6	mitotic cell division	0.777184	0.001085
LI.M8	E2F transcription factor network	0.811219	0.001153
LI.M27.1	chemokine cluster (II)	0.700175	0.001249
LI.M35.0	signaling in T cells (I)	0.741508	0.001711
LI.M61.0	enriched in NK cells (II)	0.745482	0.001711

LI.M10.1	E2F1 targets (Q4)	0.777504	0.001711
LI.M86.1	proinflammatory dendritic cell, myeloid cell response	0.686548	0.00184
LI.M71	enriched in antigen presentation (I)	0.713108	0.002373
LI.M157	enriched in NK cells (III)	0.822109	0.002373
LI.M112.0	complement activation (I)	0.687387	0.00395
LI.M15	Ran mediated mitosis	0.809206	0.004188
LI.M7.0	enriched in T cells (I)	0.637452	0.004679
LI.M95.0	enriched in antigen presentation (II)	0.696265	0.005101
LI.M27.0	chemokine cluster (I)	0.649879	0.005872
LI.M61.1	enriched in NK cells (KIR cluster)	0.952797	0.006633
LI.M4.11	mitotic cell cycle in stimulated CD4 T cells	0.815779	0.007165
LI.S5	DC surface signature	0.57073	0.009301
LI.M49	transcription regulation in cell development	0.632027	0.010549
LI.M4.10	cell cycle (II)	0.778996	0.017228
LI.M35.1	signaling in T cells (II)	0.706575	0.017815
LI.M76	DNA repair	0.693131	0.01828
LI.M4.12	C-MYC transcriptional network	0.715134	0.018418
LI.M5.0	regulation of antigen presentation and immune response	0.561311	0.01851
LI.M7.1	T cell activation (I)	0.62737	0.018856
LI.M4.9	mitotic cell cycle in stimulated CD4 T cells	0.749141	0.020322
LI.M130	enriched in G-protein coupled receptors	0.794312	0.023019
LI.M38	chemokines and receptors	0.706293	0.024886
LI.M86.0	chemokines and inflammatory molecules in myeloid cells	0.568602	0.028193
LI.M22.1	mismatch repair (II)	0.694001	0.028193
LI.S10	Resting dendritic cell surface signature	0.550361	0.029807
LI.M20	AP-1 transcription factor network	0.627408	0.033011
LI.S3	Plasma cell surface signature	0.665601	0.033898
LI.M73	enriched in monocytes (III)	0.64378	0.034145
LI.M122	enriched for cell migration	0.720921	0.038456
LI.M61.2	enriched in NK cells (receptor activation)	0.683523	0.042002
LI.M95.1	enriched in antigen presentation (III)	0.594848	0.04938

Supplementary Table 6 Module enrichment in the genes sorted by weights in PC2 and PC6 of gene expression set of TB patients

ID	Title	AUC	adj.P.Val
	Enrichment in PC2		
LI.M7.0	enriched in T cells (I)	0.866599	1.18E-13
LI.M7.1	T cell activation (I)	0.79087	9.02E-07
LI.M171	heme biosynthesis (I)	0.961933	2.04E-06
LI.M245	translation initiation factor 3 complex	0.8971	1.18E-05
LI.M222	heme biosynthesis (II)	0.921906	1.70E-05
LI.M7.2	enriched in NK cells (I)	0.772867	3.27E-05
LI.S0	T cell surface signature	0.76349	0.000118
LI.M7.3	T cell activation (II)	0.726406	0.000238
LI.M61.2	enriched in NK cells (receptor activation)	0.901331	0.000654
LI.M173	erythrocyte differentiation	0.806322	0.000997
LI.M18	T cell differentiation via ITK and PKC	0.883852	0.001108
LI.M61.0	enriched in NK cells (II)	0.836805	0.001281
LI.M7.4	T cell activation (III)	0.872297	0.002163
LI.M5.1	T cell activation and signaling	0.685075	0.003272
LI.M14	T cell differentiation	0.836531	0.003272
LI.M19	T cell differentiation (Th2)	0.728054	0.003627
LI.M212	purine nucleotide biosynthesis	0.777762	0.008759
LI.M238	respiratory electron transport chain (mitochondrion)	0.725251	0.010242
LI.M126	double positive thymocytes	0.727144	0.016192
LI.M117	cell adhesion (GO)	0.615905	0.020669
LI.M234	transcription elongation, RNA polymerase II	0.698233	0.036267
LI.M157	enriched in NK cells (III)	0.72714	0.036267
LI.M167	enriched in cell cycle	0.685136	0.049391
	Enrichment in PC6		
LI.M14	T cell differentiation	0.891986	0.002228
LI.M147	intracellular transport	0.804421	0.004291
LI.M7.4	T cell activation (III)	0.740942	0.013856
LI.M117	cell adhesion (GO)	0.724468	0.020221
LI.M144	cell cycle, ATP binding	0.688033	0.030775
LI.M18	T cell differentiation via ITK and PKC	0.750013	0.031709
LI.M109	receptors, cell migration	0.729665	0.042816
LI.M7.0	enriched in T cells (I)	0.609577	0.042816
LI.M11.0	enriched in monocytes (II)	0.585239	0.046486
LI.M230	cell cycle, mitotic phase	0.825074	0.047172